Atmospheric Environment 92 (2014) 191–198

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Point source influence on observed extreme pollution levels in a monitoring network

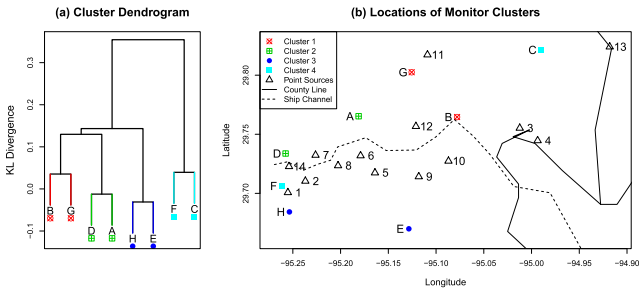
Katherine B. Ensor [a,*], Bonnie K. Ray [b], Sarah J. Charlton [a]

[a] Department of Statistics, MS 138, Rice University, Houston, TX 77251-1892, USA
[b] Business Analytics and Math Sciences, IBM T. J. Watson Research Center, USA

HIGHLIGHTS

- Benzene levels above a health-based threshold are studied for an industrial port.
- Introduce methods for count time series regression, Gaussian plume and model based clustering.
- An association between point source emitters and high benzene levels in an industrial port is found.
- On average, high ambient exposure to benzene one day is followed by high exposure the next.
- Identified a monitor anomaly of lower benzene levels than expected.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form
5 April 2014
Accepted 10 April 2014
Available online 13 April 2014

Keywords:
Extreme pollution
Point source
Count regression
Zero inflation
Model based clustering

ABSTRACT

This paper presents a strategy to quantify the influence major point sources in a region have on extreme pollution values observed at each of the monitors in the network. We focus on the number of hours in a day the levels at a monitor exceed a specified health threshold. The number of daily exceedances are modeled using observation-driven negative binomial time series regression models, allowing for a zero-inflation component to characterize the probability of no exceedances in a particular day. The spatial nature of the problem is addressed through the use of a Gaussian plume model for atmospheric dispersion computed at locations of known emissions, creating covariates that impact exceedances. In order to isolate the influence of emitters at individual monitors, we fit separate regression models to the series of counts from each monitor. We apply a final model clustering step to group monitor series that exhibit similar behavior with respect to mean, variability, and common contributors to support policy decision making. The methodology is applied to eight benzene pollution series measured at air quality monitors around the Houston ship channel, a major industrial port.

1. Introduction

Extreme pollution levels present an important public health concern, especially in cities with large industrial manufacturing and/or shipping regions, such as Houston, TX. In order to guide regulatory policy, it is important to understand the influence that major point sources in a monitoring network region have on extreme pollution values observed at each monitor. In this paper, we introduce a strategy to identify point source impact on time series observed at each monitor by modeling observed hourly counts of exceedances above a pollutant threshold. We focus our

* Corresponding author.
E-mail addresses: kbensor@comcast.net, ensor@rice.edu (K.B. Ensor).

http://dx.doi.org/10.1016/j.atmosenv.2014.04.017

study on benzene levels that exceed 0.4 parts per billion volume (ppbv). There are no national ambient air quality standards for non-criteria pollutants. In the state of Texas non-criteria pollution, such as hazardous air pollutants are monitored and compared with health effects screening levels. If a monitored level exceeds this screening level, the area is placed on a watch list (Raun, 2014). There are no specific regulatory consequences associated with watch list areas. The purpose of the Air Pollutant Watch List is to alert technical staff to cities or counties within the state that have areas with elevated air concentrations of special interest pollutants. Our consideration is on counts of high levels of pollution as a proxy for duration of exposure. First, an observation-driven negative binomial regression model is used to capture autocorrelation in daily counts over time. Because there are many days in which more zero counts (representing no exceedances) are observed than would be expected for a negative binomial distribution, we include a zero-inflation component to account for this effect. Furthermore, we use the Gaussian plume model (GPM) for atmospheric dispersion to create covariates designed to measure the impact of emissions based on the locations of the leading point source contributors. These covariates represent the effect of an emissions source registered at a monitor. We also incorporate atmospheric conditions, such as wind speed, wind direction, and solar radiation in our covariate construction. Finally, we develop a model-based approach to clustering the zero inflated count series obtained from each monitor. Understanding the common patterns in counts of observed threshold exceedances allows us to identify similarities in the influence of the point sources on sets of monitors, as well as enables us to identify monitoring sites, often spatially contiguous, representing similar (and dissimilar) behavior in pollution patterns.

The paper is organized as follows. In Section 2.1, we develop our new methodology consisting of the observation driven zero inflated negative binomial (ODZINB) time series model, creation of the point source covariates from the Gaussian plume model (GPM), and the model based clustering for time series. In Section 3 we use our developed methodology to understand extreme benzene patterns for eight series of daily counts observed for two years in the Houston area. We conclude in Section 4.

2. Material and methods

2.1. Modeling time series of counts

Time series of counts can arise in many diverse applications, for example, finance, quality control, epidemiology, and, in this case, environmental science. The ambient benzene measurements considered here exhibit large deviations and extreme values. A direct analysis of the measured levels using methods from extreme value theory is provided in Su et al. (2012). However, as noted by these authors there are outliers that are highly influential and omitted from the analysis due to their impact. By focusing on counts above a fixed threshold, we are able to understand the general pattern of the exceedances, thereby providing additional statistical methodology to understand the complex data. Further, daily counts of hourly levels of pollution above a threshold serve as a proxy for duration of exposure. Count models based on the Poisson or negative binomial (NB) distributions have been extended to the time series setting with either a parameter-driven model [e.g. Zeger, 1988], or an observation-driven model [e.g. Davis et al., 2005; Woodard et al.]. Fokianos and Kedem (2004) developed an observation-driven Poisson regression model which takes advantage of existing techniques from generalized linear modeling (GLM) to estimate model parameters. We leverage their approach here.

2.1.1. Generalized linear time series model

The basic idea is to model the transformed response variable of counts $\{Y_t,\ t = 1,\ldots,N\}$ as a linear function of the covariates $\{\boldsymbol{Z}_t,\ t = 1,\ldots N\}$, including, in the case of time series, past values of the response. The distribution of the response conditioned on the past belongs to the exponential family of distributions, capturing the random component of the GLM through

$$f(y_t;\theta_t,\phi|\mathcal{F}_{t-1}) = \exp\left\{\frac{(y_t\theta_t - b(\theta_t))}{a_t(\phi)} + c(y_t,\phi)\right\}, \tag{1}$$

where θ_t is the parameter of the distribution, ϕ is a scale parameter, and $\mathcal{F}_{t-1}$ represents the observations up to time $t-1$. The systematic component of the GLM is represented by the mean of Y_t, denoted as μ_t, modeled by a monotone link function $g(\cdot)$ such that $g(\mu_t) = \mathbf{Z}_t'\boldsymbol{\beta}$, where $\mathbf{Z}_t$ is the set of covariates for the process Y_t, and $\boldsymbol{\beta}$ is the vector of coefficients for the linear combination of the covariates. The specific form of $g(\cdot)$ depends on the underlying distribution of the count series. Both the Poisson and the NB distributions are exponential family models with common link function $g(\mu_t) = \log(\mu_t)$. Our benzene exceedance data exhibits overdispersion relative to the Poisson, motivating us to use a NB model. A negative binomial distribution specified by a mean parameter, μ, and a dispersion parameter, α has the following probability mass function.

$$P(Y = y) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!}\left(\frac{\alpha}{\alpha+\mu}\right)^{\alpha}\left(\frac{\mu}{\alpha+\mu}\right)^{y}. \tag{2}$$

In our analysis, we assume a constant dispersion parameter, α, and allow the mean, μ, to change over time as a function of covariates. While alternative approaches for capturing the overdispersion in the Poisson model could also be used, such as the recently developed Conway–Maxwell–Poisson distribution that can fit count data with a wide range of dispersion levels (Sellers and Shmueli, 2010), the NB distribution proved adequate for the purposes of our application.

Following Fokianos and Kedem (2004), a straightforward way to include serial dependence in the model is to include past observations of the response Y_t, as well as exogenous covariates in the link function model, yielding $\mathbf{Z}_t'\boldsymbol{\beta} = \beta_0 + \beta_1\log(Y_{t-1}+1) + \ldots + \beta_p\log(Y_{t-p}+1) + \beta_{p+1}X_t + \ldots + \beta_{p+q+1}X_{t-q}$, where X_t is some time dependent covariate, and p and q represent the needed time lags. The transformation $\log(Y_{t-j}+1)$ of Y_{t-j} for $j = 1,\ldots,p$ is needed to ensure the model is well behaved (Cameron and Trivedi, 1998).

A partial likelihood approach is used to estimate the parameters of the model. An advantage of the partial likelihood is that it can be obtained directly using standard GLM estimation methods, such as the `glm` function in the R statistical software (R Development Core Team, 2009). For further explanation of the partial likelihood in the count time series setting see (Kedem and Fokianos, 2002).

2.1.2. The zero-inflated count model

Data consisting of counts may sometimes have more than the expected number of zeros. If we are measuring the number of hours in a day pollution levels are above a specified health limit, *the more zeros the better*. To quantify and characterize the zero inflation in our series we use the basic methodology developed by Lambert (1992). In this model, the responses, $\mathbf{Y} = \{Y_1,\ldots,Y_N\}$ are independent with $Y_t = 0$ with probability p_t and Y_t has the distribution $F(\boldsymbol{\theta}_t)$, with probability $1-p_t$, where $\boldsymbol{\theta}$ are the parameters of that distribution. The model states that the positive counts and some of the zero counts are correctly specified by the count distribution, while extra zeros are observed according to the mixing parameter, p_t.

The parameters $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_T)'$ satisfy $\log(\boldsymbol{\mu}) = \mathbf{Z}_t'\boldsymbol{\beta}$ and $\mathbf{p} = (p_1,\ldots,p_T)'$ satisfy $\text{logit}(\mathbf{p}) = \log(\mathbf{p}/(1-\mathbf{p})) = \mathbf{G}_t'\boldsymbol{\gamma}$ for covariates $\mathbf{Z}_t$ and $\mathbf{G}_t$ and corresponding coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. In the basic definition of the zero-inflated model, the covariates $\mathbf{Z}_t$ and $\mathbf{G}_t$ may be the same, partially coincide, or be completely different (Lambert, 1992).

As in Lee et al. (2005), we modify the basic zero-inflated negative binomial regression model to incorporate functions of lagged values of the response, Y_t, as covariates in both the count and zero-inflation components of the ZINB model. Thus, we call this model the observation-driven zero-inflated negative binomial (ODZINB), with the name "observation-driven" placed first to indicate that both the zero-inflation process and the negative binomial counts are data-driven. For estimating the ODZINB models, we used methods available in the `pscl` package of the R statistical software (R Development Core Team, 2009; Zeileis et al., 2008).

2.2. Building point source covariates

Based on the location of the point sources we use the Gaussian plume model (GPM) for atmospheric dispersion [e.g. Wark et al., 1981, Ch. 4] to construct covariates representing the effect on pollution levels at the location of an air quality monitor from a major point source of benzene emissions. The goal of including these covariates in the model is to identify which point sources significantly impact the daily counts of exceedances at the air monitors.

Our approach with the GPM differs from the usual applications of the plume equation. Typically, the GPM is used for predicting the concentrations of airborne chemicals at a fixed distance downwind from an emissions source. Our use of the GPM is similar to a class of inverse atmospheric dispersion problems, where the location of unknown point source emissions are predicted based on recorded pollution levels at an air quality monitor; see e.g. (Islam and Roy, 2002; Rudd et al., 2012). In a similar application, Khemka et al. (2006) studied threshold exceedances in the inverse setting; air monitors recorded only whether an air pollution threshold was exceeded, and the researchers used their methods to predict the locations of emissions sources.

The GPM model in its full generality takes into account dispersion of air sources in longitude, latitude and vertical distance. However, in our case, we do not know the vertical height, or stack height, at which the pollution is emitted, nor the emission rate. Therefore, we make the simplifying assumption that the pollution is measured at the same height as it was emitted, and also assume an emission rate of one. It is important to note that we are creating covariates for our regression model, hence the regression coefficient will partially capture missing information such as emission rate. Under these simplifying assumptions, the GPM reduces to:

$$C(x_t, y_t) = \frac{1}{2\pi\sigma_{y_t}\sigma_{z_t}u_t}\exp\left(\frac{-y_t^2}{2\sigma_{y_t}^2}\right) \equiv S_t(x_t, y_t) \tag{3}$$

where $C(\cdot)$ represents the concentration (g/m^3) of the pollutant at location (x_t,y_t), the spatial coordinates relative to the point source, at time point t; u_t is the wind speed (m/s); and σ_{y_t} and σ_{z_t} are crosswind and vertical dispersion parameters, respectively. The dispersion parameters depend on x_t, the effective downwind distance from the point source, and the Pasquill–Gifford air stability class (Wark et al., 1981), which is a function of wind speed and degree of solar insolation (cloud cover). Standard Pasquill–Gifford tables are available in Turner (1997), and a modification for using solar radiation as a proxy for cloud cover, as we have done, is found in Katz (2006). GPM values are obtained at every hour and then averaged across the day to yield one daily value (Turner, 1997) for each monitor and point source combination. Larger values are associated with shorter distances between monitors and point sources.

2.3. Model based clustering for time series

Clustering of air monitoring sites into groups that display similar behavior of pollution levels and trends can help streamline public policy decisions regarding ambient air quality standards (Ignaccolo et al., 2008; Morlini, 2007). A few clustering techniques have been applied to pollution data from air monitor networks. One approach is to treat time series from N monitors as vectors and employ a clustering algorithm based on a Euclidean distance or Pearson's correlation distance between the series. This is the basic approach by Lavecchia et al. (1996) and Saksena et al. (2003), who then used hierarchical clustering, and also Gramsch et al. (2006), who used a condition of intra-cluster variance to partition the clusters. Others, such as Ignaccolo et al. (2008) and Morlini (2007), have taken a functional data approach to clustering air monitoring networks. In particular, Ignaccolo et al. (2008) used a partitioning around medoids algorithm to obtain a representative object for each cluster that could be used to quickly assess the nature of air pollution in each of the clusters. Morlini (2007) used a dynamic time warping cost (DTWC) to align time series of pollution concentrations and then utilized the calculated pairwise DTWC as a metric for hierarchical clustering.

In our case, with only eight monitors, the objective is to identify monitor series that behave similarly with respect to the number and pattern of exceedances and the impact of point source covariates. Ideally one would want to fit a multivariate zero inflated NB count model, however the heterogeneity observed in the series makes such a model difficult. A multivariate approach is also difficult to implement due to the differing patterns of missing values across the series. A common model within clusters provides an acceptable trade-off in terms of model simplicity and descriptive power, with the GPM-based point source covariates used to capture spatial effects. Also, by combining observations from monitors that naturally cluster, statistical power for determining point source impact is increased. As the number of series increase, the information gleaned from model based clustering has the opportunity to highlight the commonalities across the region. Analyzing the residuals from the cluster models, both within and between clusters, for spatial–temporal correlation is an important diagnostic check.

With the ODZINB model for counts we apply a model-based clustering technique to find groups of Houston air monitors having similar patterns of air pollution. We use an empirical Kullback–Leibler (KL) divergence measure (Zhong and Ghosh, 2003) to quantify the similarity or dissimilarity between the modeled time series. Let λ_j be a specified model structure for the data $\{Y_1,Y_2,\ldots,Y_N\}$. Then the KL divergence can be defined as

$$D_{\text{KL}}(\lambda_k,\lambda_j) = \frac{1}{|K|}\sum_{Y\in K}\left[\log p(Y|\lambda_k) - \log p(Y|\lambda_j)\right], \tag{4}$$

where K is the set of data objects which belong to cluster k, $p(\cdot)$ is the underlying probability function for the model, and λ_k refers to the model structure for cluster k. In our application, λ is an ODZINB regression model with specified covariates, while λ_k refers to the estimated ODZINB model for a specific cluster, k. For count time series, the KL divergence uses log of the partial likelihood as described in Section 2.1. We use normalized likelihoods for quantifying the divergence between series of varying lengths. The KL divergence measure between clusters k and j can be made

symmetric, thereby satisfying the triangle inequality for distance metrics, by defining

$$D_{\mathrm{SKL}}(\lambda_k, \lambda_j) = \frac{D_{\mathrm{KL}}(\lambda_k, \lambda_j) + D_{\mathrm{KL}}(\lambda_j, \lambda_k)}{2}. \quad (5)$$

We then use the symmetric KL divergence given by Equation (5) in a model-based hierarchical agglomerative clustering algorithm; the general algorithm is presented in Table 1. As a unique step in the clustering algorithm, we have included a test for zero-inflation. Count series are fit first with a standard count model and then tested for extra zeros. Vuong's test (Vuong, 1989) or some other goodness-of-fit test to a non-inflated count distribution can be used, and if indicated, a zero-inflated model is fit to the data. Note that all members of a cluster must have the same choice of models, that is, the same distributional assumption with or without zero-inflation.

By using a hierarchical clustering algorithm, we have minimized the need for any stopping criterion or search for an optimal number of clusters. In this application, we chose clusters by inspecting the resulting cluster dendrogram. Note the algorithm does not include any model selection devices, so once the clusters are identified, a final model selection procedure must be undertaken manually. In this application we use a simple backwards step selection.

3. Results and discussion

With the ODZINB model definition, our strategy for creating point source covariates via the GPM and the technology to cluster common time series, we proceed with the application of our newly developed tools. Our case study focuses on benzene observed at eight monitors in the Houston-area for the years 2006 and 2007. For previous work in line with this methodology focusing on 1,3Butadiene we refer the interested reader to (Thomas et al., 2009). In addition to being the fourth largest city in the United States, Houston has a major port and a large petrochemical industry, each of which provide many sources of air pollution (Sexton et al., October 2007). The volatile organic compound (VOC) benzene plays a contributing role to the rapid formation of ozone in the Houston region, but also itself poses serious health risks [see Raun et al., 2009; Sexton et al., October 2007; Whitworth et al., 2011 and references therein]. Benzene is known to be a human carcinogen (Smith et al., 2007). This is a great public concern because industrial sources in Houston produce more benzene than any other US location. Thus controlling air pollution and protecting the health of Houston's citizens is both a serious concern and a formidable challenge.

In 2006 and 2007 more than 30% of reported benzene concentrations at the eight monitors exceeded the 1×10^{-5} inhalation risk level. The screening level is equal to 0.4 parts per billion volume (ppbv) and can be interpreted as the risk of an additional person in a population of 100,000 developing cancer as a result of prolonged exposure to this level of chemical. We focus on counts of high levels of pollution as a proxy for duration of exposure. Our approach of modeling the counts of threshold exceedances as opposed to hourly averages is in line with the Texas Commission on Environmental Quality (TCEQ) goals of understanding and monitoring the long term patterns of very high benzene levels common in the region to reduce large pollution events (TCEQ, 2005a).

Table 1
General agglomerative hierarchical algorithm for clustering zero-inflated time series of counts. The model structure λ refers to an appropriate regression model for non-negative counts, with specified covariates.

Input: A set of N time series $Y = \{Y_1, \ldots Y_N\}$, and model structure λ.
Output: A cluster hierarchy of the time series.
Algorithm:

1. Initialize each data series as its own cluster and train a model for each cluster, *i.e.*, $\lambda_n = \max_\lambda \log p(Y_n|\lambda), n = 1,\ldots,N$;
2. Test series for zero-inflation and fit a zero-inflated model if necessary;
3. Compute the pairwise distance between clusters according to equation (5);
4. Merge the two closest clusters, say k and j, and re-train a model for the new cluster $k = k \cup j$, *i.e.* $\lambda_k = \max_\lambda \log p(Y_k|\lambda)$;
5. Stop if all series have been merged into one cluster, otherwise return to step 2.

3.1. Data description

Pollution concentrations and other meteorological data are available from air monitoring stations maintained by the TCEQ. As benzene monitoring is a developing program for the TCEQ, only eight monitors in the Houston metropolitan region (Harris County, TX) currently collect data on benzene. The locations of the benzene monitoring stations in Harris County are labeled A through H in Fig. 1. The eight monitors are located near the Houston Ship Channel, (TCEQ, 2005a), representing the area of highest emissions. The map in Fig. 1 marks the approximate path of the ship channel and locations of industrial facilities that are known major sources of benzene pollution.

Monitoring stations take five-minute samples of ambient air and report an hourly average of these five-minute samples. The data series, Y_{it}, is the number of hours in day t at monitor i that the benzene concentration measurements exceed the threshold. The values of Y_{it} range from 0 to n_{it}, where n_{it} denotes the number of non-missing hourly observations at site i on day t. In this data set the maximum value of n_{it} is 22, due to regularly scheduled recalibration of the monitors for two hours every night. If all the hourly observations for a day are missing, that day is omitted in the estimation of the regression model. It is conceivable that a non-truncated NB regression model might predict some counts to be above 22, the maximum number of exceedances we can observe in a day. For our series the truncation error is nominal; the median probability value of the count exceeding 22 based on our model fits is between 0.004 and 0.04 for the eight series.

3.2. Model covariates

There are two sets of covariates for the ODZINB regression model: covariates for the NB count process and covariates for the

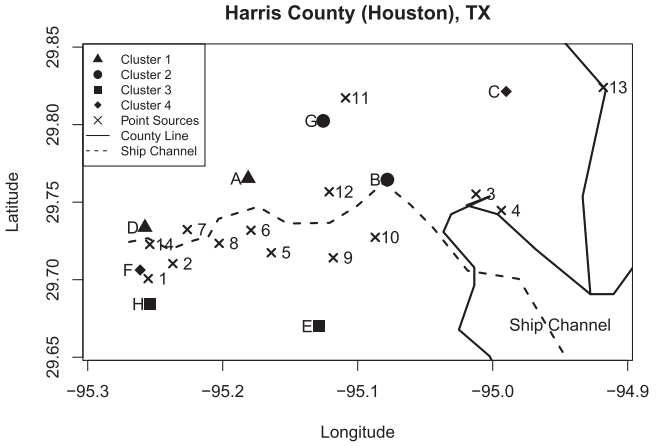


Fig. 1. Map showing location of top fourteen point sources of benzene emissions relative to the corresponding eight air monitors. Ships come in from the Gulf of Mexico.

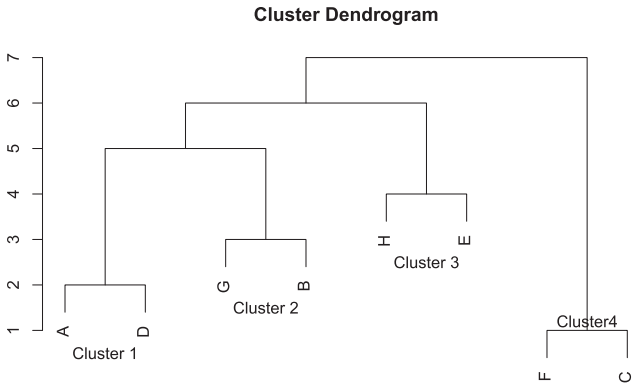


Fig. 2. Dendrogram for model based clustering.

zero-inflation process. For the NB component, in addition to the GPM covariates representing point sources of benzene emissions, we include up to two transformed lags of the observed count as covariates. We also find improved explanatory power if we include exogenous covariates wind speed, solar radiation, and information on wind direction. The covariates for modeling the zero-inflation component of the model again include seven data-driven transformed lag of count terms, as well as exogenous covariates wind speed, solar radiation, and wind direction. Wind speed (daily average) and solar radiation (third quartile) determine general atmospheric stability levels [see e.g. Wark et al., 1981]. The characteristics of the ambient air at a monitoring station effect the number of observed zeros. The wind in Harris County and Southeast Texas tends to originate from the southeast; TCEQ air quality monitors which track benzene have been positioned generally to the northwest of industrial areas in Harris County such as the Ship Channel. When the wind blows from the southeast, the monitors are more likely to record higher pollution levels. To capture this dominant wind pattern, we use the daily proportion of hours in which the wind blows from the southeast, rather than an average daily wind direction in our model.

The covariates reflect the mixture nature of our ODZINB model and the fact that there are two sources of zeros, namely zeros from the NB counts and extra zeros from zero-inflation. A monitor may record zero exceedances for a day if the effect from point source emissions is low, and this will be reflected in the estimated NB count for that day. However, a monitor may record zero exceedances for a day due to some local atmospheric conditions, regardless of the effect of point sources, and this behavior is captured as extra zeros from the ZINB mixture.

3.2.1. *Covariate construction using the Gaussian plume model*

Point sources of VOC emissions in the Houston area include facilities such as petroleum product refineries, chemical factories, and shipping including the Port of Houston which is one of the largest in the U.S. These point sources are major contributors to the benzene pollution in Harris County (Buzcu and Fraser, 2006; TCEQ, 2005a).

There are many industrial point sources in Harris County, mostly concentrated around the industrial ship channel. The TCEQ requires all these entities to self-report emissions of air toxics. Facilities are identified as the major polluters by TCEQ if they emit more than ten tons aggregate of a single VOC or more than 25 tons aggregate total VOCs for the year; we have this list of facilities based on year 2005 emission inventories, the most recently available data (TCEQ, 2005b). Based on these reports, we identified the locations of the 14 worst known benzene polluters; see the map of Harris County in Fig. 3(a), which also shows the locations of the air monitors.

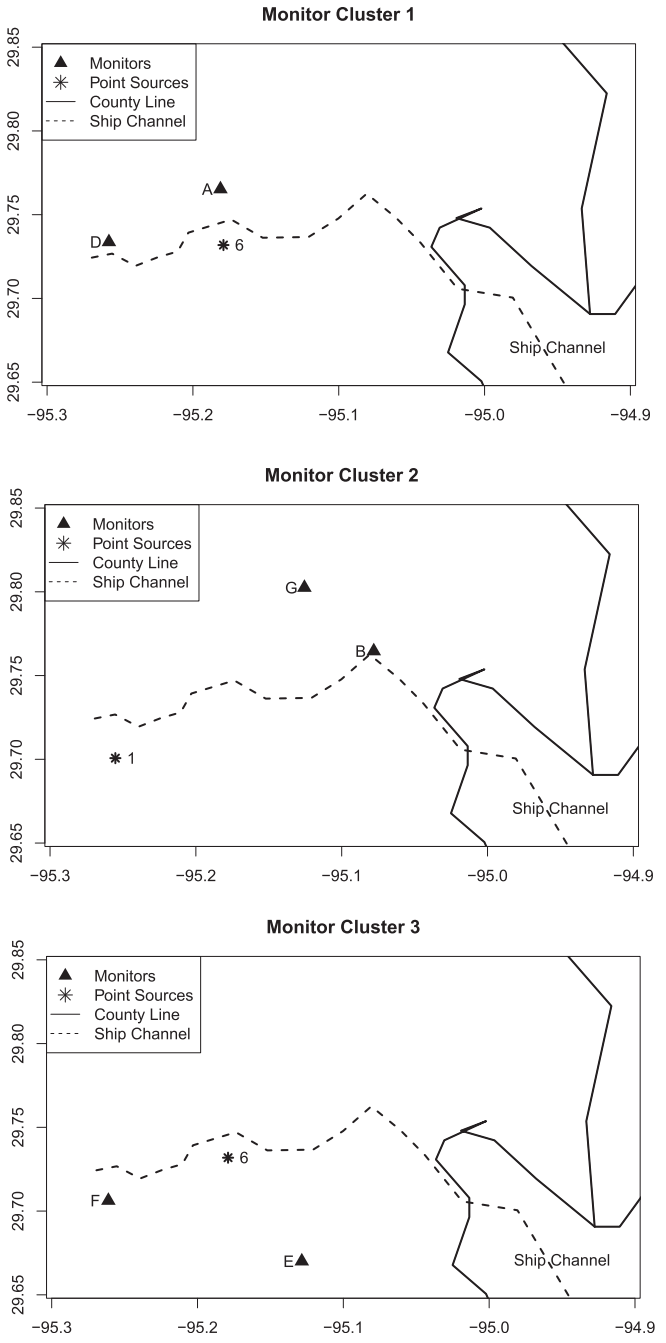


Fig. 3. Thumbprint maps showing locations of significant point sources of benzene emissions relative to the corresponding air monitors based on the ODZINB model fit for each cluster.

Using the GPM methodology, we create covariates for all pairs of monitors i and point sources j, yielding S_{ijt}, the "effect" on monitor i, from point source j, on day t. For each monitor $i = 1,\ldots,8$, we have 14 different covariates derived from the GPM as defined in Section 2.2, representing the 14 major VOC polluters in the Houston area. The covariate values range from [0,1], representing the inverse distance based on the GPM, between the monitor and the location of the point source.

3.3. *Model summary*

The final model for characterizing the $i = 8$ daily time series of extreme benzene counts, representing the eight air quality

monitors in the Houston ship channel, is an ODZINB regression model with covariates:

- S_{ijt} representing the influence from the fourteen point sources ($j = 1,\ldots,14$) derived from the Gaussian plume model to the eight monitors ($i = 1,\ldots,8$) for each day,
- $W^S_{i,t}$ is the average daily wind speed in (m/s) for each of the eight monitors,
- $R_{i,t}$ is the average daily solar radiation (Langley/min) at each of the eight monitors,
- and $W^D_{i,t}$ is the proportion of the hourly wind resulting from the SE direction throughout the day at each of the eight monitors.

The predominant wind pattern is from the southeast as the wind comes in from the Gulf of Mexico. In addition to the covariates, there is the dispersion parameter for the NB, α, which is estimated as a constant for each model.

The model has two components, namely modeling the transformed mean of the number of exceedances within a day, given by $\log(\boldsymbol{\mu}_{i,t})$, and the transformed probability of no exceedances on any given day, given by $\text{logit}(\boldsymbol{p}_{i,t}) = \boldsymbol{p}_{i,t}/1 - \boldsymbol{p}_{i,t}$. The transformations of *log* and *logit* represent the standard GLM link functions for NB counts and Bernoulli events, respectively. Both the number of exceedances and the excess probability of no exceedances are modeled as a linear function of the covariates previously described.

The mathematical summary of the model is given by:

$$\begin{aligned}\log(\boldsymbol{\mu}_{i,t}) &= \mathbf{Z}^T_{i,t}\boldsymbol{\beta};\\ \mathbf{Z}_{i,t} &\equiv \Big(1, \log(Y_{i,t-1}+1), \log(Y_{i,t-2}+1),\\ &\qquad S_{i,1,t}, S_{i,2,t}, \ldots, S_{i,14,t}, , R_{i,t}, W^D_{i,t}, W^S_{i,t}\Big)^T\\ \text{logit}\Big(\boldsymbol{p}_{i,t}\Big) &= \mathbf{G}^T_{i,t}\boldsymbol{\gamma};\\ \mathbf{G}_{i,t} &\equiv \Big(1, \log(Y_{i,t-1}+1), \log(Y_{i,t-2}+1), R_{i,t}, W^D_{i,t}, W^S_{i,t}\Big)^T\end{aligned} \tag{6}$$

where $Y_{i,t}$ is the count series of the number of hours in a day that exceeded the threshold for monitor i and day t, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the respective regression coefficients.

The eight time series of benzene counts, which represent eight air quality monitors around Houston, were fit with the ODZINB model described above. Due to power considerations when modeling individual series, we allowed variables to remain in the model if the associated p-value was 0.1 or less. We tighten the modeling in the second stage of the paper after identifying monitor clusters. For each series, we also applied Vuong's test for non-nested models (Vuong, 1989) to test for the presence of zero-inflation. The zero-inflated model was preferred in all cases.

3.4. Results of model based clustering

Model based clustering was then performed on the eight models according to algorithm 1. The cluster dendrogram can been seen in Fig. 2. For discussion, we name the clusters one through four, from the left to right of Fig. 2. Thus cluster 1 contains monitors A and D, cluster 2 has monitors B and G, cluster 3 has monitors E and H, and cluster 4 contains the remaining monitors C and F. Within each cluster, the best ODZINB model was obtained. The final model results are given in Table 3. We have also included Table 2 which provides summary information for each cluster as well as the mean absolute error and square root of the mean squared error between the observed and predicted counts. A pseudo-R^2 value is obtained by squaring the Pearson correlation between the observed and predicted counts and shows our models explain between 30% and 44% of the variation in the data. Further, the residuals from the cluster based models demonstrate no significant serial or spatial correlation either within or between clusters, based on Spearman's rank based correlation measure.

Table 2
Average daily hours exceeding the threshold, percent of days with no exceedances and model diagnostics by cluster, and for all eight series combined (ALL). Model diagnostics include: Mean Absolute Error (MAE), the square root of the Mean Squared Error (MSE), and a pseudo-R^2 computing by squaring Pearson correlation between the observed and predicted counts from the ODZINB model.

Monitor characteristics and model fits					
Cluster (monitors)	1 (B, G)	2 (A, D)	3 (E, H)	4 (C, F)	ALL monitors
Average count	10.4	8.5	7.0	4.1	7.5
% Days count = 0	3	8	11	31	13
MAE	3.85	3.79	3.26	2.73	3.81
Square root of MSE	4.79	4.75	4.21	3.67	4.82
Pseudo-R^2	0.30	0.35	0.44	0.35	0.34

3.4.1. Interpretation of model based clusters

The three clusters which contain monitors A and D, B and G, and E and H, respectively, are grouped based on spatial closeness. Cluster 1 is on the northern side of the industrial ship channel area, while cluster 2 is to the east/northeast, and cluster 3 is south of the ship channel. This geographical clustering is a consequence of the GPM-based covariates we employed for characterizing the monitors, and is a feature we were seeking in our model development. Similar autocorrelation in the benzene counts, similar effects from point sources based on our GPM covariate construction, and similar behavior of zero-inflation are all features we would expect geographically close air monitors to share. The only non-geographic pair consists of monitors C and F (Cluster 4, see Fig. 1) which appear to be clustered together because they have lower counts overall compared to the other monitors.

3.4.2. General interpretation of the zero-inflation

Overall, the results from fitting the ODZINB model are consistent with the nature of the benzene pollution in the region. We start with the zero-inflation component, because it should be interpreted similarly across all the individual models for each monitor. Estimated regression coefficients for the zero-inflation part of the model are presented in the lower half of Table 3. The large negative

Table 3
Estimated ODZINB models for benzene count series for each of the four clusters of air quality monitors and all monitors combined. Covariates with an * are significant at the 0.05 level; all other covariates are significant at the 0.01 level. Empty spaces indicate the covariate was not significant at the 0.05 level.

Cluster (monitors)	1 (A, D)	2 (B, G)	3 (E, H)	4 (C, F)	All monitors
Covariate	Coefficients with p-value < 0.01				
Negative binomial count process					
Intercept	2.04	1.98	2.03	2.37	1.79
Point sources	$S_6 = 0.51$	$S^*_1 = 0.40$	$S_6 = 0.62$	–	$S_6 = 0.37$
$\log(Y_{t-1}+1)$	0.12	0.25	0.31	0.29	0.33
R_t	−0.81	−0.18	−0.67	−0.57	−0.46
W^D_t	0.53	−0.47	−0.80	−0.56	−0.25
W^S_t	−0.25	−0.16	−0.31	−0.36	−0.20
$\log(\alpha)$	1.82	2.12	1.77	1.12	1.46
Zero-inflation process					
Intercept	−3.01	−4.95	−3.58	−3.26	−2.80
$\log(Y_{t-1}+1)$	−0.82	−0.65	−0.71	−0.55	−1.01
R_t	−3.43	–	–	–	–
W^D_t	−1.60	1.04*	1.51	1.64	0.85
W^S_t	1.03	0.93	0.65	0.93	0.81

intercept terms indicate a low baseline of zero-inflation implying most of the probability for extra zeros arise from the effect and behavior of the covariates in the model.

For all models, we have significant negative coefficients for the lag one term. The implication is that if a high count is observed one day, we are less likely to observe a zero count due to zero-inflation the next day. This result mirrors the observed positive autocorrelation in the negative binomial part of the count model (see next sub-section), and shows the usefulness of our approach for including a data-driven component in the zero-inflation part of the ODZINB model.

Wind speed has positive estimated coefficients in all cases; the positive coefficient implies that as wind speed increases, the probability of zero-inflation increases. This agrees with the science of atmospheric dispersion; increasing wind speed carries any pollutants in the air downwind more rapidly, in a sense cleaning out the air in the region and increasing the probability of zero exceedances. The estimated coefficients for wind speed have similar values across models, as can be seen in Table 3, indicating that wind speed affects the probability of zero-inflation similarly for the eight monitors.

Wind direction does not have uniform signs on estimated model coefficients. For clusters 2, 3 and 4, the wind direction covariate has a positive effect on the probability of zero-inflation. Recalling that the wind direction covariate, W_t^D, is defined as the proportion of hours per day in which the wind originates from the southeast, the effect of wind direction on zero-inflation can have opposite signs depending on where monitors are located relative to point sources. In the map of Harris County benzene monitors (Fig. 3), we see that monitors E and H are on the southern edge of the area covered by the monitors, and monitor C is to the far east. Thus, the major point sources of benzene pollution are located to the north of monitors E and H. So when wind blows from the southeast the major benzene emissions are blown away from these monitors and we are more likely to have zero-inflation. In contrast, monitors A and D are located exactly in the path of benzene emissions flowing to the northwest, thus decreasing the probability of zero-inflation.

The fact that cluster one (monitors A and D) exhibit negative correlation between zero-inflation probability and solar radiation is also explained in terms of atmospheric properties. Increases in solar radiation represent increased dispersion and more air movement. Sites A and D sit in close proximity directly downwind from major point sources, thus increased air movement results in lower probability of zero-inflation. Solar radiation was not a significant contributor in explaining the probability of zero exceedances for the other clusters at the level of 0.05 level.

3.4.3. General interpretation of the count regression

Now we consider the general interpretation of the NB count regression component of the model. All of the estimated models include the lag of the count series, $\log(Y_{t-1}+1)$, as a significant covariate with positive coefficients, indicating a residual level of benzene pollution still lingering in the ambient air from the previous day. Recall that the "autocorrelation" terms in the zero-inflation part of the model were negative, so for high counts, not only are we more likely to observe a high count the next day, but also less likely to observe a zero.

Point source covariate S_6 appears in clusters one, two and three as well as in the model when combining all monitors. Point source covariate S_1 is a significant variable for cluster one at the 0.05 level of significance. Our models indicate that emissions from these point sources increase the number of hours in a day that benzene will exceed our set threshold at the associated monitors.

There are no significant point source covariates in cluster four which includes monitors C and F. Monitor F is in close proximity to and directly downwind from the number one emitter in the region. However, the level of benzene observed at monitor F is low and exhibits the same pattern as site C, a naturally cleaner site far removed from the major benzene point sources. Further, the number one emitter is not a significant point source for monitor F (or C). Such an unusual statistical pairing suggests additional investigation of monitor F to identify the factors leading to the unexpected low levels of benzene observed. For example, it may be that the stack height of the largest benzene point source is far above the monitoring level, thereby mitigating the impact of the benzene emissions on the nearby monitors.

To assess the fit of individual monitors models, we take the estimated model coefficients and compute the estimated zero-inflated negative binomial mean, ζ, as

$$E\left[Y|\widehat{\beta},\widehat{\alpha},\widehat{\gamma}\right] = \widehat{\zeta}_t = (1-\widehat{p}_t)\widehat{\mu}_t, \tag{7}$$

where $(\widehat{\beta},\widehat{\alpha},\widehat{\gamma})$ are the estimated model parameters and $\widehat{\mu}_t$ and $\widehat{p}_t$ are the resulting NB mean and zero-inflation processes, respectively, based on those coefficients in Equation (6). The residuals are computed in the usual manner, $r_t = Y_t - \widehat{\zeta}_t$. Examination of the residuals indicated good model fits and lack of residual autocorrelation. Goodness of fit measures are provided in Table 2

The resulting estimated models for each cluster are presented in Table 3. Overall, we have positive dependence on past counts in the count component of the model and negative dependence on past counts in the zero-inflation part of the model. This indicates that a large number of daily exceedances is persistent in time. This finding is of concern as it is an indication of prolonged exposure by residents in the area. A significant positive coefficient for the GPM covariate indicates the point source is contributing to the high levels of benzene observed at the monitor. The primary contributors to elevated counts above the set threshold are point sources one and six. Again, this contribution is characterized based on the meteorological adjusted distance the point source is from the monitor where the patterns are observed at a daily level over a two year period.

Examining the meteorological variables, we see that as expected wind speed is an important variable for each cluster model. Cluster 1 has a negative relationship between zero-inflation and solar radiation, indicating that as air stability increases, represented by a decrease in solar radiation, we are more likely to observe excessive levels of pollution at sites A and D, which are in close proximity to major emitters.

4. Conclusions

In this paper, we have developed and applied a new strategy for investigating leading point sources affecting observed extremes in air quality monitoring data as measured at a network of air quality monitoring stations. The strategy is based on modeling measured concentrations above a specified health threshold using an observation-driven zero-inflated count regression model coupled with model based clustering. In our application to extreme benzene concentrations measured in and around the Houston, Texas ship channel, covariates for the regression model were derived from the Gaussian plume equation for atmospheric dispersion, representing a transformed distance from a point source of benzene emissions to the air monitoring station. Our results indicate that point source covariates are important factors for modeling the counts of daily benzene measurements that exceed a given health risk threshold. Improvements would be expected if we extended our Gaussian plume modeling to include vertical height. Coupled with other studies of benzene pollution in the Houston region, these insights

can be used to evaluate the effectiveness of local air quality management. Further, the development of count time series regression models and their clustering may aid in further understanding of public health related studies where the number of exceedances in a day is used as a proxy for duration of ambient exposure to benzene.

The methodologies presented work with any pollutant where daily levels above a threshold are of interest. As previously noted in the paper, earlier versions of this work addressed 1,3butadiene as well as benzene. We presented only the benzene results to bring clarity and focus to the manuscript.

Acknowledgments

The authors acknowledge the generous financial support of the National Science Foundation, United States (grants DMS-0240058 and DMS-0739420) and Houston Endowment, United States. Further, the authors acknowledge the air quality scientists for the City of Houston, and specifically Dr. Loren Raun, who have been generous with their time and expertise throughout our research. We would also like to thank the reviewers for their insightful comments. The manuscript is much improved as a result of their efforts.

References

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press.

Buzcu, B., Fraser, M.P., 2006. Source identification and apportionment of volatile organic compounds in Houston, TX. Atmospheric Environment 40 (13), 2385–2400.

Davis, R.A., Dunsmuir, W.T.M., Streett, S.B., 2005. Maximum likelihood estimation for an observation driven model for Poisson counts. Methodology and Computing in Applied Probability 7 (2), 149–159.

Fokianos, K., Kedem, B., 2004. Partial likelihood inference for time series following generalized linear models. Journal of Time Series Analysis 25 (2), 173–197.

Gramsch, E., Cereceda-Balic, F., Oyola, P., von Baer, D., 2006. Examination of pollution trends in Santiago de Chile with cluster analysis of PM_{10} and ozone data. Atmospheric Environment 40 (28), 5464–5475.

Ignaccolo, R., Ghigo, S., Giovenali, E., 2008. Analysis of air quality monitoring networks by functional clustering. Environmetrics 19 (7), 672–686.

Islam, M., Roy, G., 2002. A mathematical model in locating an unknown emission source. Water Air and Soil Pollution 136 (1–4), 331–345.

Katz, D.I., October 2006. The scoop on sensor selection. Environmental Protection Magazine 17 (8), 36–49. http://www.eponline.com/articles/54198/.

Kedem, B., Fokianos, K., 2002. Regression Models for Time Series Analysis. Wiley-Interscience.

Khemka, A., Bouman, C.A., Bell, M.R., 2006. Inverse problems in atmospheric dispersion with randomly scattered sensors. Digital Signal Processing 16.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1–14.

Lavecchia, C., Angelino, E., Bedogni, M., Brevetti, E., Gualdi, R., Lanzani, G., Musitelli, A., Valentini, M., 1996. The ozone patterns in the aerological basin of Milan (Italy). Environmental Software 11 (1–3), 73–80.

Lee, A.H., Wang, K., Yau, K.K., Carrivick, P.J., Stevenson, M.R., 2005. Modelling bivariate count series with excess zeros. Mathematical Biosciences 196 (2), 226–237.

Morlini, I., 2007. Searching for structure in measurements of air pollutant concentration. Environmetrics 18 (8), 823–840.

R Development Core Team, 2009. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raun, L., April 3 2014. Personal Communication.

Raun, L.H., Marks, E.M., Ensor, K.B., 2009. Detecting improvement in ambient air toxics: an application to ambient benzene measurements in Houston, Texas. Atmospheric Environment 43 (20), 3259–3266.

Rudd, A., Robins, A., Lepley, J., Belcher, S., 2012. An inverse method for determining source characteristics for emergency response applications. Boundary-Layer Meteorology 144 (1), 1–20.

Saksena, S., Joshi, V., Patil, R.S., 2003. Cluster analysis of Delhi's ambient air quality data. Journal of Environmental Monitoring 5, 491–499.

Sellers, K.F., Shmueli, G., 2010. A flexible regression model for count data. Annals of Applied Statistics 4 (2), 943–961.

Sexton, K., Linder, S.H., Marko, D., Bethel, H., Lupo, P.J., October 2007. Comparative assessment of air pollution related health risks in Houston. Environmental Health Perspectives 115 (10), 1388–1393.

Smith, M.T., Jones, R.M., Smith, A.H., 2007. Benzene exposure and risk of non-hodgkin lymphoma. Cancer Epidemiology Biomarkers & Prevention 16 (3), 385–391.

Su, F.-C., Jia, C., Batterman, S., 2012. Extreme value analyses of VOC exposures and risks: a comparison of RIOPA and NHANES datasets. Atmospheric Environment 62 (0), 97–106.

TCEQ, Spring 2005a. On the Track of Air Pollution: Agency Fine-tunes Monitoring Project Along Houston Ship Channel.

TCEQ, 2005b. Emissions Inventory. Unpublished Report. Texas Commission on Environmental Quality.

Thomas, S., Ray, B., Ensor, K.B., 2009. A Model-based Approach for Clustering Air Quality Monitors in Houston, Texas. Tech. Rep. TR2009-04. Rice University, Department of Statistics.

Turner, D.B., 1997. The long lifetime of dispersion methods of Pasquill in U.S. regulatory air modeling. Journal of Applied Meteorology 36, 1016–1020.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 (2), 307–333.

Wark, K., Warner, C.F., 1981. Air Pollution: its Origin and Control, second ed. HarperCollins, New York.

Whitworth, K.W., Symanski, E., Lai, D., Coker, A.L., 2011. Kriged and modeled ambient air levels of benzene in an urban environment: an exposure assessment study. Environmental Health 10, 21.

Woodard, D.B., Matteson, D.S., Henderson, S.G., 2011. Stationarity of generalized autoregressive moving average models. Electronic Journal of Statistics 5, 800–828. http://dx.doi.org/10.1214/11-EJS627. http://projecteuclid.org/euclid.ejs/1312818919.

Zeger, S.L., 1988. A regression model for time series of counts. Biometrika 75 (4), 621–629.

Zeileis, A., Kleiber, C., Jackson, S., 2008. Regression models for count data in R. Journal of Statistical Software 27 (8), 1–25.

Zhong, S., Ghosh, J., 2003. A unified framework for model-based clustering. Journal of Machine Learning Research 4, 1001–1037.