Journal of Environmental Statistics
August 2013, Volume 5, Issue 5. http://www.jenvstat.org

Use of the Dagum Distribution for Modeling Tropospheric Ozone levels

Benjamin Sexto M.
Colegio de Postgraduados

Humberto Vaquera H.
Colegio de Postgraduados

Barry C. Arnold
UC, Riverside

Abstract

This paper deals with the use of the Dagum distribution to model the maximum daily levels of tropospheric ozone. We compare the fit of the Dagum distribution against the Generalized Extreme Value distribution (GEV) by using the Kolmogorov-Smirnov test and the Akaike criterion for model selection. Also we propose a methodology for estimating long term trends in the daily maxima of tropospheric ozone by using the Vector Generalized Linear Model (VGLM) and quantiles of the Dagum distribution. Ozone data from Pedregal Station in Mexico City (one with the worst air pollution in the World) are analyzed for the period 2001-2008. Results show that the Dagum model has a similar or better fit than the GEV model. The quantiles of Dagum distribution and VGLM show evidence of a downward trend in high ozone levels at Pedregal Station.

Keywords: trends, urban ozone, extreme value, vgam.

1. Introduction

One important pollutant in big cities is ozone (O_3) which in high levels (above .12 ppm) is harmful to human health (Ebi and McGregor 2008). Urban ozone effects may be more severe in certain susceptible groups such as children, elderly, sick people and people who enjoy outdoor exercise (Ponce de Leon, Anderson, Bland, and Bower 1996).

In tropospheric ozone data analysis the traditional distributions used are the Generalized Extreme Value distribution (GEV) and the Pareto distribution.

Extreme values in environmental time series are important because of their applicability to the analysis of catastrophic phenomena such as extreme ozone observations, and extreme meteorological conditions (floods, winds, temperature, etc). The statistics of extremes can undoubtedly be useful in applications relating to distributions with light or bounded tails, but they are found to be most useful for variables that have a heavy tailed distribution (Katz,

Parlange, and Naveau 2002).

The Dagum distribution has two parameters, one of shape and the other of scale. This distribution has been used by economists as a distribution for modeling country incomes because of its property of having a heavy right tail. Mielke (1973) used the Kappa distribution (with three parameters) to model the amount of rainfall precipitation. The Kappa distribution Mielke and Johnson (1974) includes the Dagum distribution in a different parametrization (referred as the Beta-K distribution). Dagum (1977) and Fattorini and Lemmi (1979) proposed the Kappa distribution as an income distribution.

In this paper use of the Dagum distribution is proposed for modeling daily maximum levels of ozone at a specific location. Subsequently, the fit of the Dagum distribution and that of the generalized extreme value distribution (GEV) are compared. An additional goal is to propose methodology for estimating long term trends in the daily maxima of tropospheric ozone, using information from the environmental monitoring station in Pedregal, Mexico City.

1.1. Dagum distribution

The Dagum distribution is a heavy-tailed distribution developed by Camilo Dagum in the 70's for modeling income distributions as an alternative to the Pareto (Pareto 1895) and log-normal (Gibrat 1931) models. The most general form of the Dagum distribution has the following cumulative distribution function.:

$$F(x) = \alpha + (1-\alpha)[1 + (x/b)^{-a}]^{-p} \tag{1}$$

The Dagum distributions of Type I, II and III correspond to cases where $\alpha = 0$, $0 < \alpha < 1$ and $\alpha < 0$ respectively. The Dagum type II distribution was proposed as a model for income distribution allowing for zero or negative income. It seems especially appropriate for wealth data, where there are often a large number of economic units with zero net assets. The Dagum distribution of Type III is associated with a positive lower limit for X, x_0. In this paper we will work with the Dagum of type I. Henceforth this distribution will be simply referred as the Dagum distribution. The Dagum distribution is a special case of the generalized beta distribution of the second kind (GB2). The density of the GB2 distribution is:

$$f(x) = \frac{ax^{ap-1}}{b^{ap}B(p,q)\,[1 + (x/b)^a]^{p+q}}, \quad x > 0 \tag{2}$$

where $b > 0$ is the scale parameter and $a > 0$, $p > 0$, $q > 0$ are the shape parameters. In (2), if the shape parameter q is set equal to 1, the Dagum density is obtained:

$$f(x) = \frac{apx^{ap-1}}{b^{ap}\left[1 + \left(\frac{x}{b}\right)^a\right]^{p+1}}, \quad x > 0 \tag{3}$$

where $a, b, p > 0$. The Dagum distribution function has a closed form:

$$F(x) = \left[1 + \left(\frac{x}{b}\right)^{-a}\right]^{-p}, \quad x > 0 \tag{4}$$

where a, b, $p > 0$. The parameter b is the scale parameter, while a and p are shape parameters.

In the case in which $ap > 1$ the density has an interior mode. The mode for Dagum distribution is:

$$x_{mode} = b\left(\frac{ap-1}{a+1}\right)^{1/a} \tag{5}$$

The quantile function also has a closed form:

$$F^{-1}(u) = b\left[u^{-1/p} - 1\right]^{-1/a}, \quad for \;\; 0 < u < 1 \tag{6}$$

The k-th moment exists for $-ap < k < a$ as follows:

$$E(X^k) = \frac{b^k \Gamma\left(p + k/a\right) \Gamma\left(1 - k/a\right)}{\Gamma(p)} \tag{7}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

In particular the mean and variance are:

$$E(X) = \frac{b\Gamma\left(p + 1/a\right) \Gamma\left(1 - 1/a\right)}{\Gamma(p)} \tag{8}$$

$$var(X) = \frac{b^2\left[\Gamma(p)\Gamma\left(p + 2/a\right) \Gamma\left(1 - 2/a\right) - \Gamma^2\left(p + 1/a\right) \Gamma^2\left(1 - 1/a\right)\right]}{\Gamma^2(p)} \tag{9}$$

In practical situations the estimated value of parameter a is usually small (in economic applications a gets smaller as income inequality increases) (Dagum and Lemmi 1989).

Parameter estimation can be implemented using the method of maximum likelihood. Let $X_1, ..., X_n$ be a random sample of size n from the Dagum distribution, the log-likelihood function is defined as:

$$\ell = n\log a + n\log p + (ap-1)\sum_{i=1}^{n}\log x_i - nap\log b - (p+1)\sum_{i=1}^{n}\log\left[1 + \left(\frac{x_i}{b}\right)^a\right] \tag{10}$$

A variety of standard optimization programs can be used to maximize this function. In particular, the package EVIR in R can be utilized.

1.2. GEV distribution

Three types of extreme value limit distributions play a fundamental role in the analysis of extremes of environmental data: Fréchet, Weibull and Gumbel. The Generalized Extreme Value (GEV) is a combination of these three types of extreme value limit distributions (von Mises (1936, 1954) and Jenkinson (1955)), and its distribution function is:

$$G(x;\mu,\sigma,\xi) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\} \tag{11}$$

where $\sigma > 0$, $-\infty < \mu < \infty$, $1 + \xi(x-\mu)/\sigma > 0$, $x_+ = max\{x, 0\}$. The parameter μ is a location parameter, σ is a scale parameter, and ξ shape parameter. The cases in which $\xi > 0$, $\xi < 0$ and $\xi = 0$, correspond to the Fréchet, Weibull and Gumbel distributions respectively. The quantile function of the GEV distribution is:

$$G^{-1}(u) = \mu - \frac{\sigma}{\xi}\left[1 - \{-\log u\}^{-\xi}\right] \tag{12}$$

with $0 < u < 1$. The value $G^{-1}(1-u)$ is the return level associated with the return period $1/u$.

2. Statistical Methodology

2.1. Ozone Data

The pollutant concentrations to be studied correspond to an urban site located South of Mexico City. These measurements are integrated in the Air Quality Monitoring Network of the Valle de Mexico Metropolitan Area, managed by the Atmospheric Monitoring System (SIMAT) of the Mexico City Government. The analyzed data correspond to daily maxima of ozone measures (ppm). Ozone concentrations were monitored using UV absorption photometry using the API 400 and API 400A. The study data correspond to the period from 2001 to 2008. The data is available at `http://www.sma.df.gob.mx/simat2/informaciontecnica`.

2.2. Block Maxima

For the ozone data set, a block length of three days is considered to be a long enough period of separation between observations to achieve independence. In each block, the maxima was obtained (Block maxima). The Block maxima method is described by (Gaines and Denny 1993). The distribution of maximum ozone levels is not the same each year; there is a trend towards lower peak levels of ozone over the years, thus the series is not stationary. Therefore, it is appropriate to analyze and adjust the maximum levels of ozone for each year to minimize the non-stationarity problem. A time series plot for block maxima of ozone levels is in Fig.(1). Autocorrelation in ozone data would have little effect on the bias of parameter estimates, but their variance is affected Vaquera H (1997). The use of block maxima reduces the undesirable consequences of the autocorrelation.

2.3. Parameter estimation for GEV and Dagum

In the case of Dagum distribution, the Parameter values for $(\hat{a}, \hat{b}, \hat{p})$ that maximize the log-likelihood were obtained using a computational routine in the "VGAM" library for R. The calculation of estimates of the parameters of the GEV model was implemented using the maximum likelihood method with the *EVIR* package for R.

Smith (1985) observed that, for the GEV distribution (11) in the case in which $\xi < -1/2$, the usual asymptotic distributional properties of maximum likelihood estimators (MLE) do not hold. In contrast, in the case of the Dagum distribution the maximum likelihood estimators do not have such problems according to Kleiber and Kotz (2003). Consequently, if the two models, GEV and Dagum, provide comparable fits to a given data set, an argument can be advanced in favor of using the Dagum model. As we shall see, this is the case for the Ozone data analyzed here.

2.4. Assessing the of Fit of the Ozone data

The Kolmogorov-Smirnov statistic was used for comparing the Dagum and GEV distribution fits of the ozone maxima time series for each year in the range 2001-2008. The test statistic

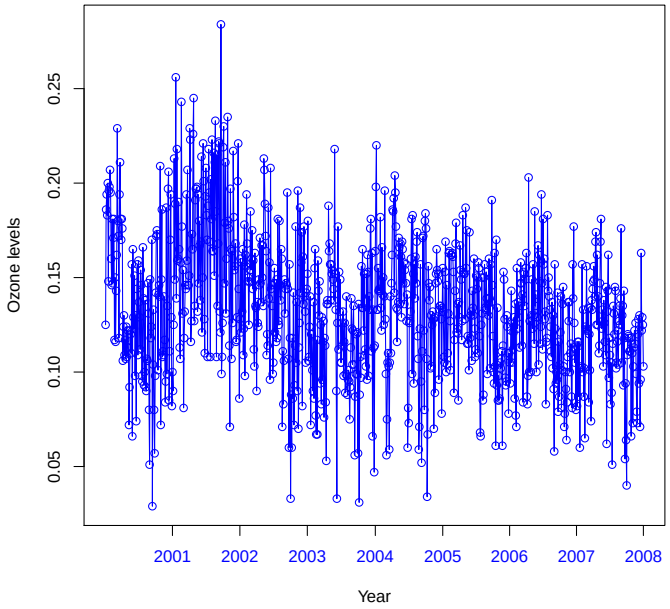

Figure 1: Daily maxima ozone levels for Pedregal (ppm)

is:

$$D = \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

where $F_n(\cdot)$ is the empirical distribution function and $F_0(\cdot)$ is the fitted distribution function with parameters estimated by maximum likelihood.

Another criteria for assessing the fit is the Akaike Information Citerion (AIC) Akaike (1974). When comparing models using the maximum likelihood method for fitting, the AIC is calculated using the following expression:

$$AIC(k) = 2k - 2log(L(\hat{\theta})) \tag{13}$$

where, k is the number of model parameters estimated by the method of maximum likelihood and $L(\hat{\theta})$ is the likelihood function evaluated at the maximum likelihood estimate $\hat{\theta}$. The preferred model will be the one with the lowest AIC.

2.5. Trend estimation in the Ozone data levels

Quantile estimates

Reyes, Vaquera, and Villaseñor (2010) proposes a statistical methodology to analyze the trends of very high values of tropospheric ozone, the methodology is based on the estimation of percentiles of the distribution of extreme values (GEV). In this work a similar idea is used for investigating trends for Dagum and GEV distribution. For the calculation of the quantile estimates, we can use maximum likelihood method with the **EVIR** package in R.

Vector Generalized Linear Model

The Vector Generalized Linear Model (VGLM) allows us to determine if there is a linear relationship between the parameters of the Dagum distribution with time as a covariate. The estimation of the parameters of the VGLM model can be performed in the VGAM library of R using the $VGLM()$ function. The VGLM and VGAM were introduced by Yee and Hastie (2003) and Yee and Wild (1996).

The VGAM/VGLM are implemented in the package VGAM (Yee 2007), working in R. The VGAM and VGLM allow all parameters of the distribution be modeled as linear or smoothed functions of covariates. Suppose the observed response y is a q-dimensional vector. The VGLM is defined as a model for which the conditional distribution of Y given the explanatory variables x is of the form:

$$f(y|x;B) = h(y, \eta_1, ..., \eta_M) \tag{14}$$

for some known function $h(\cdot)$, where $B = (\beta_1, \beta_2, ..., \beta_M)$ is a $p \times M$ matrix of unknown regression coefficients, and the j-th liner predictor is:

$$\eta_j = \eta_j(x) = \beta_j^T x = \beta_{(j)1}x_1 + \cdots + \beta_{(j)p}x_p = \sum_{k=1}^{p} \beta_{(j)k}x_k, \;\; j = 1, ..., M \tag{15}$$

where $x = (x_1, ...x_p)^T$ with $x_1 = 1$ if there is an intercept. In our case the covariate is time.

The VGAM provide extensions to VGLM additive models, the equation predictor is generalized to a sum of smoothed functions of the individual covariates:

$$\eta_j(x) = \beta_{(j)1} + f_{(j)2}(x_2) + \cdots + f_{(j)p}(x_p) = \beta_{(j)1} + \sum_{k=2}^{p} f_{(j)k}(x_k), \;\; j = 1, ..., M \tag{16}$$

The η_j are referred to as additive predictors. $f_k = (f_{(1)k}(x_k), ..., f_{(M)k}(x_k))$ is focused on uniqueness, and are estimated simultaneously using "vector smoothers". VGLM are usually estimated by maximum likelihood using Fisher scoring or Newton-Raphson.

3. Results for Mexico City Ozone levels

The visual comparison of the adjustment of the Dagum distribution with the empirical distribution and the corresponding adjustment of the GEV distribution for maximum daily ozone levels per year is found in Figure 2. From table 1 we can observe a very similar fit for the Dagum and GEV models, and in particular in the years 2003, 2004, 2005 and 2008 the Dagum fit appears to be somewhat better.

In table 1, the p-values for the Kolmogorov-Smirnov test and the Akaike values (AIC) are shown. We can conclude that maximum daily ozone observations are satisfactorily modeled by both the Dagum distribution and the Generalized Extreme Value distribution. Table 1 shows that Dagum distribution fits better in 2003, 2005 and 2008 according to both the Kolmogorov and Akaike criteria.The Overall AIC in all years(2001-2008) for Dagum was -3616.972, and for GEV -3613.

In figure(1), a non stationary pattern in the Ozone series is clear. For this reason, an analysis has been implemented for each year separately. In table 2, the analyses for the years indicates that there is a perceptible trend in the behavior of the estimated parameters a and b of the Dagum model.

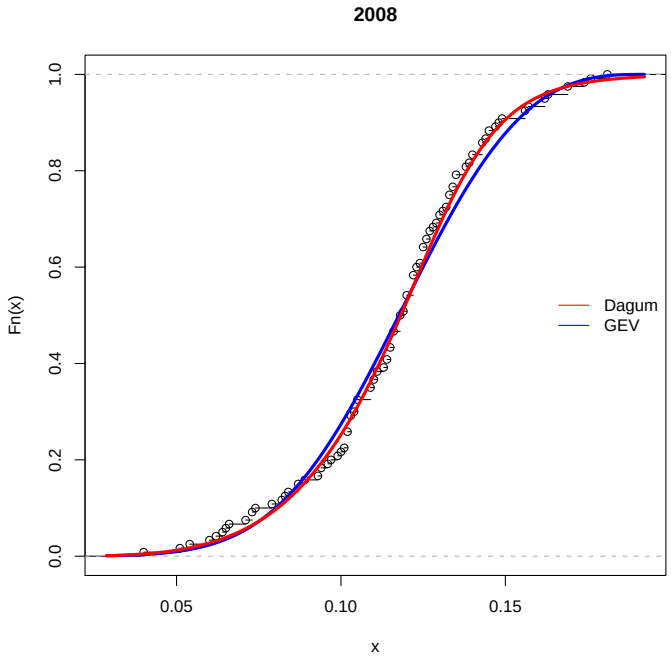

Figure 2: Example Fit Dagum vs GEV:2008

Table 1: Kolmogorov $p-values$ and AIC Akaike statistic

Year	$p-value$		AIC	
	Dagum	GEV	Dagum	GEV
2001	0.6371	0.7933	-426.6164	-434.7164
2002	0.5414	0.7623	-409.4166	-416.752
2003	0.9728	0.801	-475.4264	-474.6734
2004	0.9545	0.9743	-478.3866	-476.224
2005	0.6972	0.2813	-460.532	-456.9844
2006	0.5107	0.6088	-501.6258	-508.6648
2007	0.8596	0.9089	-502.811	-505.7758
2008	0.9318	0.5637	-518.5954	-516.0032

Table 2: Estimated parameters of Dagum Distribution

Year	Parameter		
	a	b	p
2001	8.3479	0.1615	0.4202
2002	9.7786	0.1959	0.4150
2003	13.5001	0.1669	0.2918
2004	11.2898	0.1486	0.2999
2005	14.1866	0.1701	0.2371
2006	14.9753	0.1563	0.2679
2007	10.3362	0.1394	0.4608
2008	12.3889	0.1369	0.3520

The observed change in the parameters over the years is in accordance with the results ob-

tained using a vector generalized linear model (VGLM) in which year is a covariable as follows:

$$\begin{aligned} \log(a) &= \eta_1 = \beta_{(1)1}x_1 + \beta_{(1)2}x_2 \\ \log(b) &= \eta_2 = \beta_{(2)1}x_1 + \beta_{(2)2}x_2 \\ \log(p) &= \eta_3 = \beta_{(3)1}x_1 + \beta_{(3)2}x_2 \end{aligned}$$

where $x_1 = 1$ corresponding to the intercept, x_2 is the time (years), $q = 1$ and $M = 3$. The results are presented in table 3.

Table 3: Dagum regression coefficients

Coefficients	Value	Std. Error	t-value
(Intercept):1	2.087559	0.1351299	15.44854
(Intercept):2	-1.676371	0.0367075	-45.66831
(Intercept):3	-0.920553	0.2083267	-4.41879
year:a	0.06091	0.0271134	2.24648
year:b	-0.036772	0.0065516	-5.61265
year:p	-0.022878	0.041165	-0.55575

Considering a significance level $\alpha = 0.05$ from table 3, we observe a significant linear trend in the estimated coefficients a and b (year:1 and year:2). Also the signs of these regression coefficients are consistent with the estimates from table 2. In the case of parameter p there is no significant linear trend.

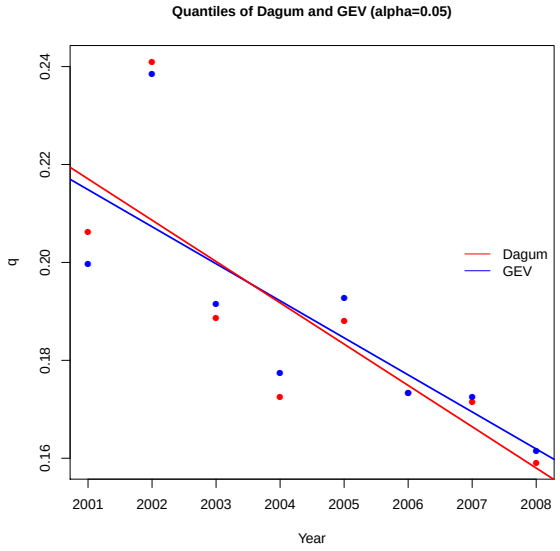

Figure 3: $(1-\alpha)100$ quantiles of Dagum and GEV

Table 4 shows the $(1-\alpha)100$ quantiles of Dagum and GEV distributions for each of the years with α=0.05 and 0.10. A downward linear trend is thus observed in these quantiles for high ozone levels.

A graphic representation is given in figure 3.

To test the proposed models for investigating trends in ozone levels, Dagum and GEV models were fitted for the period 2001-2006 and were used to forecast quantiles for the following two

Table 4: quantiles $(1-\alpha)100$ of Dagum

Year	quantiles of Dagum		quantiles of GEV	
	0.05	0.10	0.05	0.10
2001	0.2062722	0.1877332	0.1996686	0.1865772
2002	0.2410239	0.2223855	0.2384297	0.2226471
2003	0.188601	0.177531	0.1914159	0.181634
2004	0.1724811	0.1604845	0.1774077	0.1651994
2005	0.1880305	0.1772189	0.192768	0.182186
2006	0.1734096	0.1641078	0.1733466	0.1648492
2007	0.1714448	0.1589807	0.1724634	0.1611932
2008	0.1590109	0.1490738	0.1615318	0.1529513

years. The forecasted 0.95 quantiles for 2007 and 2008 were (0.167, 0.161) respectively for the Dagum model and (0.190, 0.186) in same years for the GEV model. These forecasted quantiles are generally in agreement with the fitted quantiles in figure 3 (corresponding to the years 2001-2008).

In general, note that, as expected since both models fit the data well, the difference is small between the quantiles of both Dagum and GEV distributions, and they exhibit the same trend.

4. Conclusions

With regard to the implementation of the Dagum distribution to model extreme values in ozone levels, we can conclude that:

- Based on the Kolmogorov statistic and the Akaike criteria, the Dagum distribution provides similar and sometimes better fits than does the GEV distribution for the Pedregal ozone data.
- With the results obtained in this paper, we justify the implementation of the Dagum distribution to model extreme values. The Dagum distribution is an appealing option for modeling extreme events, when using maximum likelihood estimators since it does not have the distributional problems associated with the MLE's in the GEV model.

In addition, a downward trend with time was observed for maximum ozone levels and was confirmed by two relevant techniques:

- With the Vector Generalized Linear Model (VGLM), a trend was confirmed in the estimated parameters a and b of the Dagum distribution
- The estimated $(1-\alpha)100$ quantiles corresponding to both of the Dagum and GEV models, exhibited a very similar downward trend as a function of time (years).

References

Akaike H (1974). "A new look at the statistical model indentification." *IEEE Transactions on Automatic Control*, **19**, 716–722.

Dagum C (1977). "A New Model for Personal Income Distribution: Specification and Estimation." *Economie Appliquée*, **30**, 413–437.

Dagum C, Lemmi A (1989). "A contribution to the analysis of income distribution and income inequality, and a case study: Italy." *Research on Economic Inequality*, **1**, 123–157.

Ebi KL, McGregor G (2008). "Climate Change, Tropospheric Ozone and Particulate Matter, and Health Impacts." *Environmental Health Perspectives*, **116-11**, 1449–1455.

Fattorini L, Lemmi A (1979). "Proposta di un modello alternativo per l'analisi della distribuzione personale del reddito." *Atti Giornate di Lavoro AIRO*, **28**, 89–117.

Gaines SD, Denny MW (1993). "The largest, smallest, highest, lowest, longest, and shortest: extremes in ecology." *Ecology*, **74**, 1677–1692.

Gibrat R (1931). *Les Inégalités Économiques*. Librairie du Recueil Sirey, Paris.

Jenkinson A (1955). "The frequency distribution of the annual maximum (or minimum) values of meteorological elements." *Quart. J. Roy. Meteo. Soc.*, **81**, 158–171.

Katz RW, Parlange MB, Naveau P (2002). "Statistics of extremes in hydrology." *Advances in Water Resources*, **25**, 1287–1304.

Kleiber C, Kotz S (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley, Hoboken, New Jersey.

Mielke PW (1973). "Another family of distributions for describing and analyzing precipitation data." *Journal of Applied Meteorology*, **12**, 275–280.

Mielke PW, Johnson ES (1974). "Some generalized beta distributions of the second kind having desirable application features in hydrology and meteorology." *Water Resources Research*, **10**, 223–226.

Pareto V (1895). "La Legge della Domanda." *Giornale degli Economisti, English Translation in Rivista di Politica Economica*, **10, 87(1997)**, 59–68, 691–700.

Ponce de Leon A, Anderson H, Bland J, Bower J (1996). "Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92." *J Epidemiol Comm Health*, **Vol. 50 (Supplement 1)**, S63–S70.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reyes J, Vaquera H, Villaseñor J (2010). "Estimation of trends in high urban ozone levels using the quantiles of (GEV)." *Environmetrics*, **21**(5), 470–481.

Smith RL (1985). "Maximum likelihood estimation in a class of non-regular cases." *Biometrika*, **72**, 67–90.

Vaquera H H (1997). *On the statistical analysis of trend in tropospheric ozone levels*. Ph.D. thesis, Tulane University.

von Mises R (1936). "La distribution de la plus grande de n valeurs." *Revue Mathématique de l'Union Interbalkanique (Athens)*, **1**, 141–160.

von Mises R (1954). "La distribution de la plus grande de n valeurs." *American Mathematical Society*, **Selected Papers Volumen II**, 271–294.

Yee T (2007). *A Usert's Guide to the vgam Package*. URL http://www.stat.auckland.ac.nz/~yee/VGAM.

Yee T, Hastie T (2003). "Reduced-rank Vector Generalized Linear Models." *Statistical Modelling*, **3(1)**, 15–41.

Yee T, Wild C (1996). "Vector Generalized Additive Models." *Journal of the Royal Statistical Society B*, **58(3)**, 481–493.

Affiliation:

Benjamin Sexto Monroy
Department of Statistics, Colegio de Postgraduados
Campus Montecillo,Texcoco, Mexico 56230
E-mail: bsexto@colpos.mx
URL: http://www.colpos.mx

Humberto Vaquera Huerta
Department of Statistics, Colegio de Postgraduados,
Campus Montecillo, Texcoco, Mexico 56230
E-mail: hvaquera@colpos.mx
URL: http://www.colpos.mx

Barry C. Arnold
Department of Statistics, University of California, Riverside
Riverside, CA 92521
E-mail: barry.arnold@ucr.edu
URL: http://statistics.ucr.edu/