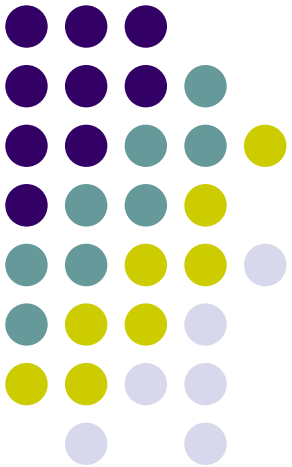
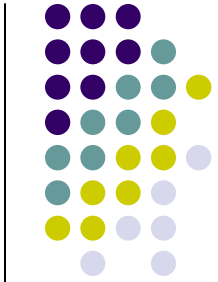
Stochastic Air Quality Analysis

Vijay P. Singh
Department of Civil and Environmental Engineering
Louisiana State University
Baton Rouge, LA 70803-6405, USA



Introduction



- Good ozone

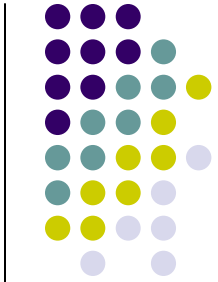
  Naturally found in the Earth's upper atmosphere 6 to 30 miles above the Earth's surface

- Bad ozone

  Found at the ground level which is considered as one of the six common air pollutants.

Ozone=NOx+VOC+Sunlight (EPA)

Introduction (cont.)



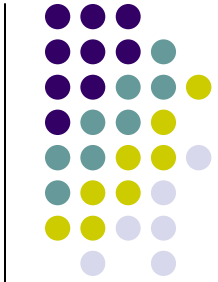
- Interested ozone variables
  - Number of days the ozone concentration exceeding the national ambient air quality standard and design value in a given year
  - Highest ozone reading in a given year
  - Duration of an ozone exceedance
  - Time interval between ozone exceedances
  - Trend over time of non-attainment parishes.

Introduction (cont.)

Air Quality Index for Ozone

Air Quality	Air Quality Index	Health protection
Good	0-50	No health impacts expected in this range
Moderate	51-100	Unusually sensitive people need consider limiting prolonged outdoor exertion
Unhealthy for sensitive groups	101-150	Active children and adults and people with respiratory diseases, e.g., asthma, need to limit prolonged outdoor exertion
Unhealthy	151-200	Active children and adults, and people with respiratory diseases, e.g., asthma, need to avoid prolonged outdoor exertion, everyone else needs to limit prolonged outdoor exertion.
Very unhealthy	201-300	Active children and adults, and people with respiratory diseases, e.g., asthma, need to avoid all outdoor exertion; everyone else, especially children, needs to limit outdoor exertion.

Copula Concept



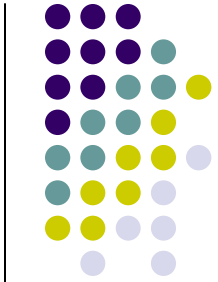
- Two random variables: *X* and *Y*
  - Observations of *X*: $x_1, x_2, \ldots, x_n$ (Non-normal population)
  - Observations of *Y*: $y_1, y_2, \ldots, y_n$ (Non-normal population)
  - CDF of *X*: $F_X(x)$
  - CDF of *Y*: $F_Y(y)$

- Let C: Dependence function-Copula

$$C(F_X(x), F_Y(y)) = H(x, y)$$

H: Bivariate distribution of *X* and *Y.*

Copula Concept (cont.)



- Let $u = F_X(x)$: marginal distribution of *X*.
  $v = F_Y(y)$: marginal distribution of *Y*.

- Bivariate distribution:

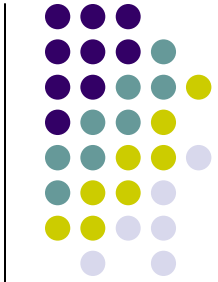
$$H_{X,Y}(x,y) = H(F_X^{-1}(u), F_Y^{-1}(v)) = C(F_X(x), F_Y(y))$$

- Archimedean copula family (Genest and Mackay, 1986)

$$C(u,v) = \varphi^{-1}(\varphi(u) + \varphi(v))$$

$\varphi(\bullet)$: generating function of C which is a continuous strictly decreasing convex function in $[0, +\infty]$ with $\varphi(1) = 0$.

Copula Concept (cont.)



- Example:

  Let *X*~exp(a); *Y*~exp(b); the copula function *C* given by the Gumbel-Hougaard copula family:

$$u = F_X(x) = 1 - \exp(x/a) \qquad v = F_Y(y) = 1 - \exp(y/b)$$

  Generating function: $\varphi(t) = (-\ln(t))^{\theta}$, $t = u$ or v

  then:

$$\varphi(u) = (-\ln(u))^{\theta} = (-\ln(1 - \exp(x/a)))^{\theta}$$

$$\varphi(v) = (-\ln(v))^{\theta} = (-\ln(1 - \exp(y/b)))^{\theta}$$

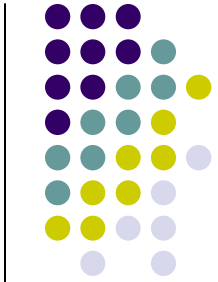
$$\varphi^{-1}(t) = \exp(-t)^{1/\theta}$$

  Thus, copula *C* can be expressed as:

$$C(u,v) = H(x,y) = \varphi^{-1}(\varphi(u) + \varphi(v))$$

$$= \exp(-((-\ln(1 - \exp(-x/a)))^{\theta} + (-\ln(1 - \exp(-y/b)))^{\theta})^{1/\theta})$$

One Parameter Archimedean Copula



- Gumbel-Hougaard copula family

$$C(u,v) = \exp\left(-\left((-\ln u)^{\theta} + (-\ln v)^{\theta}\right)^{1/\theta}\right), \theta \in [1,\infty)$$

$$\varphi(t) = (-\ln t)^{\theta}, \tau = 1 - \theta^{-1}$$

- Ali-Mikhail-Haq copula family

$$C(u,v) = \frac{uv}{1-\theta(1-u)(1-v)}, \theta \in [-1,1)$$

$$\varphi(t) = \ln\frac{1-\theta(1-t)}{t}, \tau = \left(\frac{3\theta-2}{\theta}\right) - \frac{2}{3}\left(1-\frac{1}{\theta}\right)^{2}\ln(1-\theta)$$

One Parameter Archimedean Copula (cont.)

- Frank copula family

$$C(u,v) == \frac{1}{\theta}\ln\left[1+\frac{(\exp(\theta u)-1)(\exp(\theta v)-1)}{\exp(\theta)-1}\right], \theta \neq 0$$

$$\varphi(t) = \ln\left[\frac{\exp(\theta t)-1}{\exp(\theta)-1}\right], \tau = 1-\frac{4}{\theta}\left[D_1(-\theta)-1\right]$$

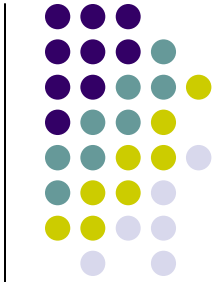
where D_1 is the first order Debye function; K-th order Debye function is defined as:

$$D_k(\theta) = \frac{k}{x^k}\int_0^\theta \frac{t^k}{\exp(t)-1}dt, \ \theta > 0$$

The k-th order Debye function with negative argument:

$$D_k(-\theta) = D_k(\theta) + \frac{k\theta}{k+1}$$

One Parameter Archimedean Copula (cont.)



- Cook-Johnson copula family

$$C(u,v)=[u^{-\theta}+v^{-\theta}-1]^{-1/\theta},\theta\geq 0$$

$$\varphi(t) = t^{-\theta} - 1, \tau = \frac{\theta}{\theta + 2}$$

Note:

- For all four copula families: $\varphi(\bullet)$ copula generating function.
- τ : Kendall's tau, which can be calculated as:

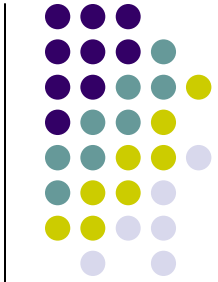
$$\tau_n = \binom{n}{2}^{-1} \sum_{i<j} sign[(x_i - x_j)(y_i - y_j)]$$

where n is the number of observations

sign=1 if $(x_i - x_j)(y_i - y_j) > 0$

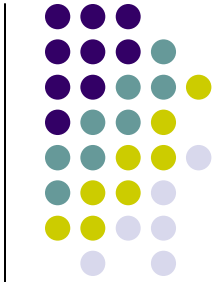
sign=-1 if $(x_i - x_j)(y_i - y_j) < 0$

Copula Identification



- Copula identification includes:
  - Calculation of Kendall's τ
  - Selection of generating function (i.e., Gumbel-Hougaard, Ali-Mikhail-Haq, Frank, Cook-Johnson etc.)
  - Estimation of copula parameter according to the relationship between Kendall's τ and generating function
  - Generating function is thus obtained
  - Copula function is therefore obtained

Selection of Best-Fitted Copula



- Selection of best fitted copula includes:
  - Intermediate random variable Z: $Z = F(X,Y)$
    - Can be obtained through

$$z_i = \{\text{No. of } (x_i, y_i) \text{ such that } x_j < x_i, y_j < y_i\}/(n-1); \text{ for } i, j = 1, 2, ..., n$$

  - Probability distribution of Z:

$$K(z) = P(Z \le z)$$

    - Nonparametric estimation of $K(z)$:

$$K_n(z) = \text{proportion of } z_i\text{'s} < z$$

    - Parametric estimation of $K(z)$ through its relationship with the copula generating function:

$$K(z) = z - \frac{\varphi(z)}{\varphi'(z)}$$

  - Q-Q plot of nonparametric and parametric estimated values of $K(z)$
  - Satisfactory copula should be the one where the plot is close to straight line passing through origin with 45 degree

Conditional Distribution Based on the Copula Method

- The conditional joint distribution between random variables *X* and *Y* can be expressed through copula as:

$$F_{X,Y}(x \mid Y = y) = P(U \le F_X^{-1}(x) \mid V = F_Y^{-1}(y)) = P(U \le u \mid V = v)$$

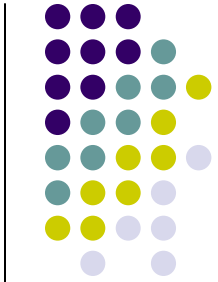
$$= \lim_{\Delta v \to 0} \frac{C(u, v + \Delta v) - C(u, v)}{\Delta v} = \frac{\partial}{\partial v} C(u, v) \mid V = v$$

and

$$F_{X,Y}(x \mid Y \le y) = P(U \le F_X^{-1}(x) \mid V \le F_Y^{-1}(y)) = P(U \le u \mid V \le v)$$

$$= C(U \le u \mid V \le v) = \frac{C(u, v)}{v}$$

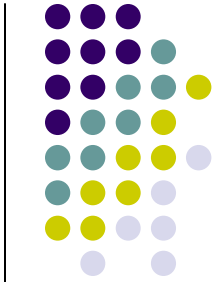
Application



- Application of Copula Method involves:
  - Determination of empirical marginal distributions
  - Identification of generator function and parameter of copulas
  - Determination of the joint probability distribution
  - Risk assessment of Ozone

- Data Description
  - Baton Rouge (1980-2000)
    - Highest ozone reading
    - Number of days of ozone violations

Application (cont.)



- Empirical Marginal Distribution
  - Plotting position formula (Gringorten, 1963; Cunnane, 1978)

$$P(K \le k) = \frac{k - 0.44}{N + 0.12}$$

where *k*: k-th smallest observation in the dataset; *N*: sample size.

- Empirical Joint Distribution
  - Simlar to the plotting position formula for marginal distribution as:

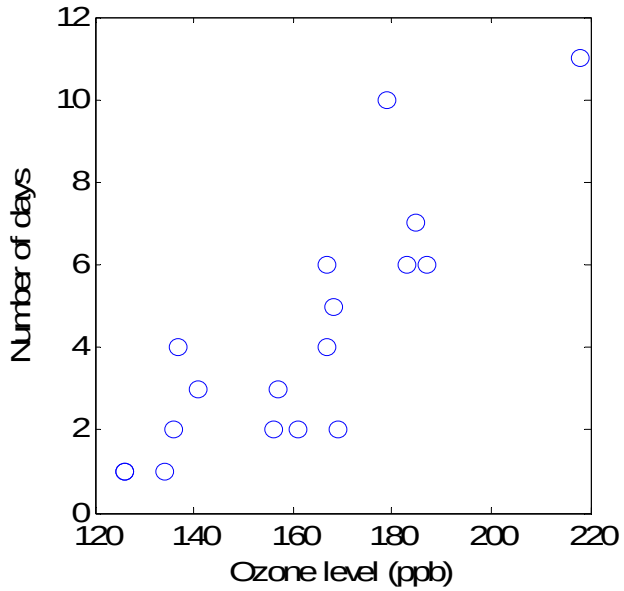
$$H(x,y) = P(X \le x_i, Y \le y_i) = \frac{\sum_{m=1}^{i}\sum_{l=1}^{i} N_{ml} - 0.44}{N + 0.12}$$

where *N* is the sample size, N_{ml} is the number of (x_j, y_j) counted

$x_j \le x_i \; and \; y_j \le y_i, i = 1,...,N$

Application (cont.)

- Dependence structure between Ozone level and number of days of violation
  - Kendall's $\tau = 0.67$
  - Pearson's product-moment correlation coefficient $\rho = 0.84$



Application (cont.)

- Joint Distribution Represented by Copula

Parameter estimated and KS-statistics for each copula candidate

	Gumbel-Hougaard	Cook-Johnson	Frank
Estimated Parameter	3	4	10.03
P-Value	0.94	0.99	0.94
KS statistics	0.17	0.11	0.17

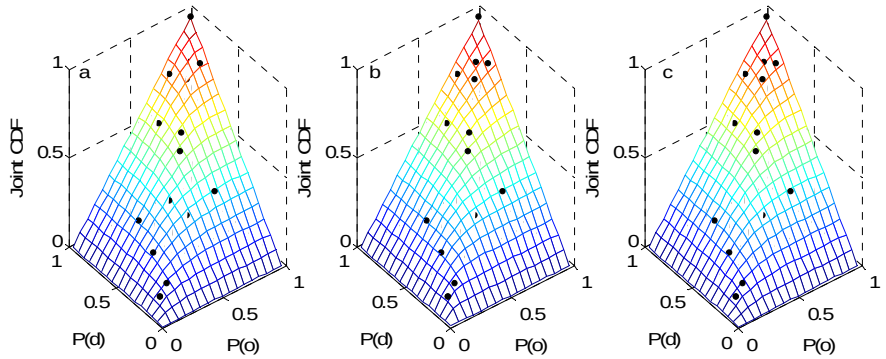


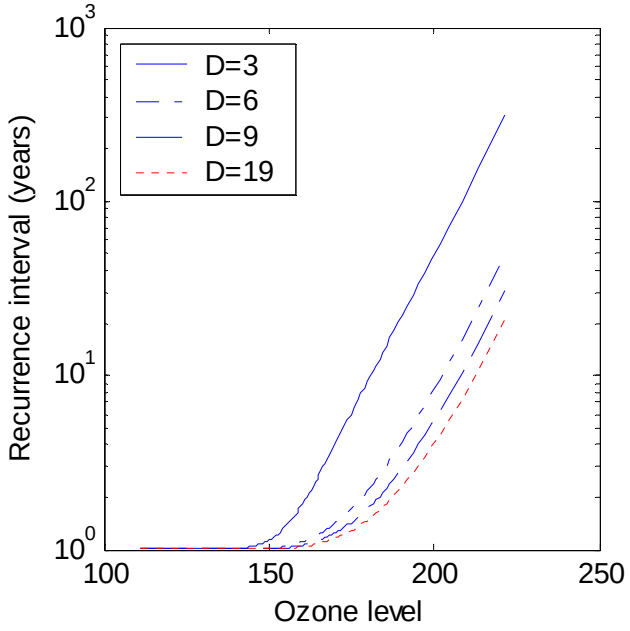
Figure: Observed and copula based probability plots.
(P(o): Cumulative probability of ozone level)
(P(d): Cumulative probability of number of days of ozone violation)
(a: Gumbel-Hougaard copula; b: Cook-Johnson copula; c: Frank copula)

Application (cont.)

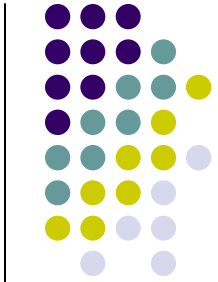
- Risk Assessment

The risk of high level ozone is represented by the recurrence interval of high ozone violations, the total number of days of ozone violation, and high ozone violations conditioned on the days of ozone violation.

Recurrence interval (years)	Ozone violation index	# days of violation	Ozone violation index conditioning variable: # of days of violation			
			D=3	D=6	D=9	D=19
2	159.78	3	161.45	178.11	183.21	187.53
5	180.54	6	172.63	193.15	198.72	203.17
10	192.06	9	180.8	202.57	208.08	212.31
20	201.94	11	189.18	210.96	215.84	220.44
50	213.45	16	200.34	230.42	250.08	260.93
100	221.35	19	208.56	283.92	326.74	346.49



Conclusions



- Considering the ozone data in Baton Rouge, the highest ozone reading and the number of days of violations are positively correlated with the Kendall correlation coefficient: 0.67 and Pearson's correlation coefficient: 0.84.
- All three copulas, i.e., Gumbel-Hougaard, Cook-Johnson, and Frank copulas are appropriate for the representation of the joint distribution of the ozone variables studied.
- Among the three copulas, it is hard to detect the most appropriate copula graphically. But according to the KS goodness-of-fit statistic, the Cook-Johnson copula is found numerically to be the most appropriate copula.
- From the conditional probability analysis it is seen that more frequently the ozone level is higher than the threshold value in a given year, more likely the highest ozone level is higher than that in a less frequently occurring year.
- According to the risk assessment simply by the recurrence interval value, people living in Baton Rouge, Louisiana, are exposed to a high risk of ozone violation.