US EPA ARCHIVE DOCUMENT

Multivariate Extremes Sub-Project: Understanding Causes of Extreme Events

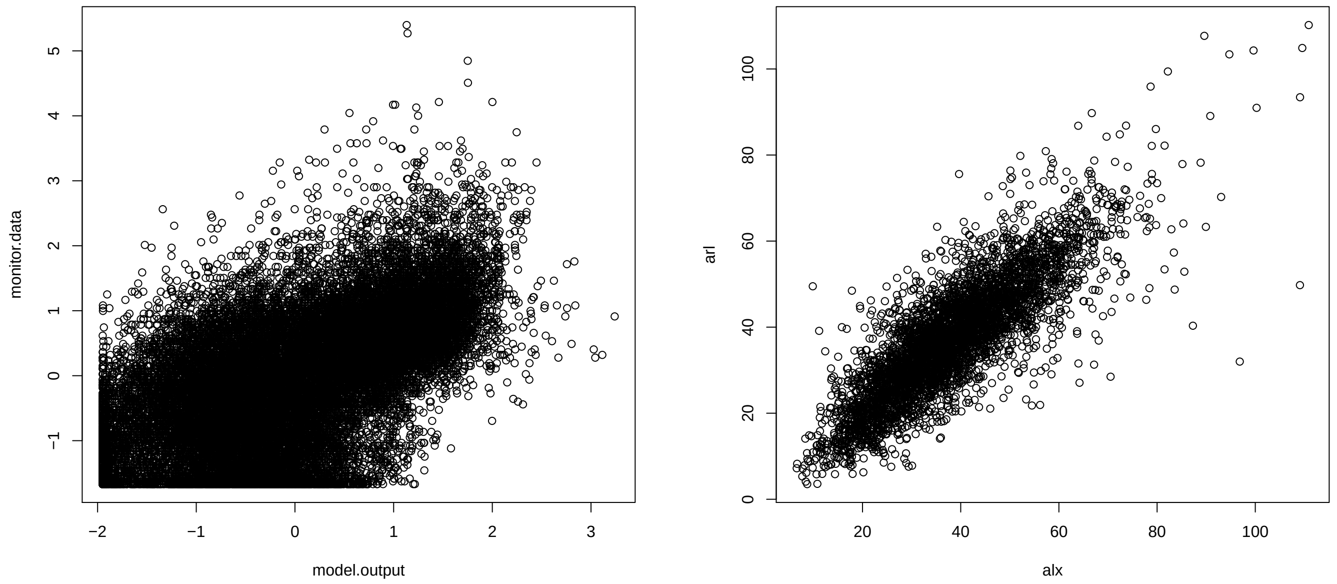
Goal: Develop *statistical methodology* to perform data mining for extreme events.

Relation to project: Can we develop statistical techniques which will help us to understand meteorological (and possibly other) factors which combine to form the 'perfect storm' resulting in the most extreme air quality events.

Approach: Techniques we develop have their basis in *extreme value theory* and specifically draw on the idea of *tail dependence*.

Acknowledgement: Brook Russell, PhD Student, CSU Statistics.

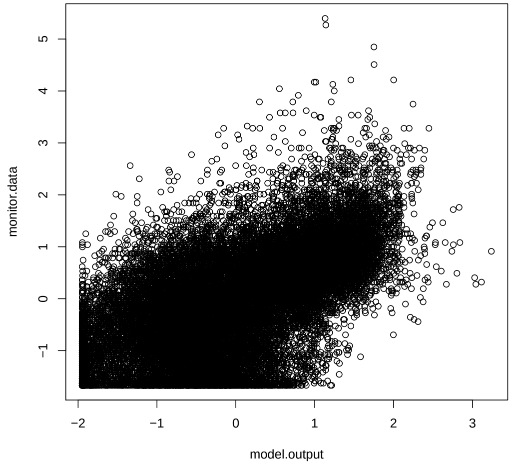
Dependence and Tail Dependence

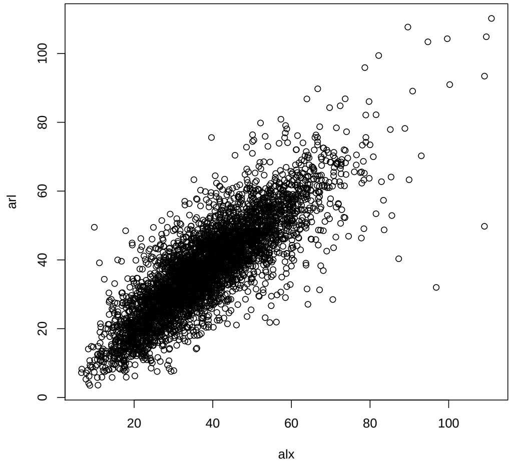


- Extreme value analyses only use extreme data.
- Multivariate extremes aims to describe tail dependence.
- Extremes a young branch of statistics, MV methodologies being developed.

NOT Tail Dependence: Correlation

$$\rho = \frac{E[(X - \mu_x)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_x)^2]E[(Y - \mu_y)^2]}}$$



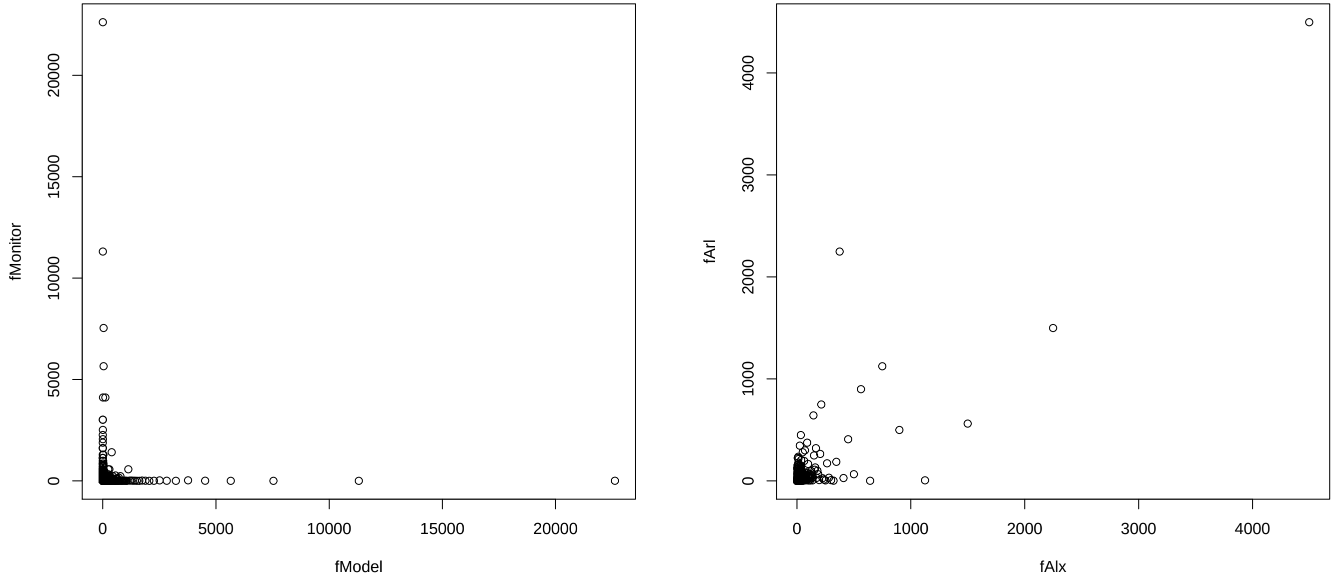


$\hat{\rho} = 0.59$ $\hat{\rho} = 0.83$

Correlation measures "spread from center", does not focus on extremes.

Viewing Tail Dependence

A technique often used to examine tail dependence is to transform to a heavy-tailed marginal distribution.



Advantages:

- Eye is drawn to large observations.
- Theoretical framework: multivariate regular variation.

Measuring Tail Dependence

There are metrics which can be used to measure tail dependence.

We will use γ, which is a parameter describing tail dependence for a regular-varying random vector.

Estimator:

$$\widehat{\gamma}(\boldsymbol{x}, \boldsymbol{y}) = n^{-1} \sum_{t=1}^{n} \frac{|x_t - y_t|}{x_t + y_t} I_{(t \in T)},$$

where T is some set corresponding to large values.

- $\widehat{\gamma}$ can be understood as measuring observed points' distance from the 45^o line.
- Small $\widehat{\gamma}$ indicates high dependence.
- $\widehat{\gamma}_1 = .761$
  $\widehat{\gamma}_2 = .501$

Data Mining Procedure via an Analogy

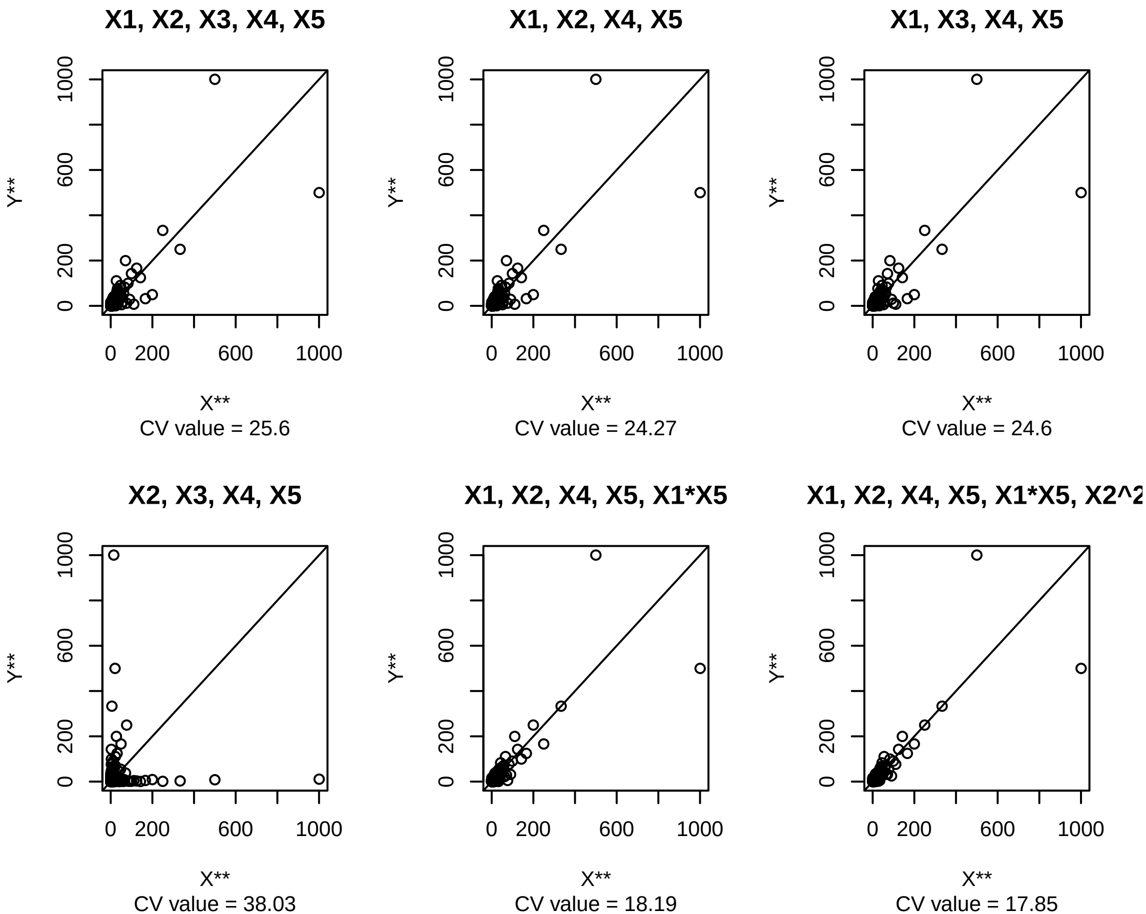
Standard Regression: find the coefficients $\boldsymbol{\beta}$ which optimize the correlation between $X_t^T\beta$ and the response Y_t. As correlation is most meaningful when both the covariates and response are normal, one may first transform the data.

Extremes: find the coefficients $\boldsymbol{\beta}$ which optimize $\widehat{\gamma}$ between $X_t^T\boldsymbol{\beta}$ and the *large values* of the response Y_t. To fit into the regular variation framework, both the response and the linear combination $X_t^T\boldsymbol{\beta}$ must have a common regularly varying marginal, thus a transformation is made.

Data Mining: A model selection problem. Which combination of covariates lead to most extreme events?

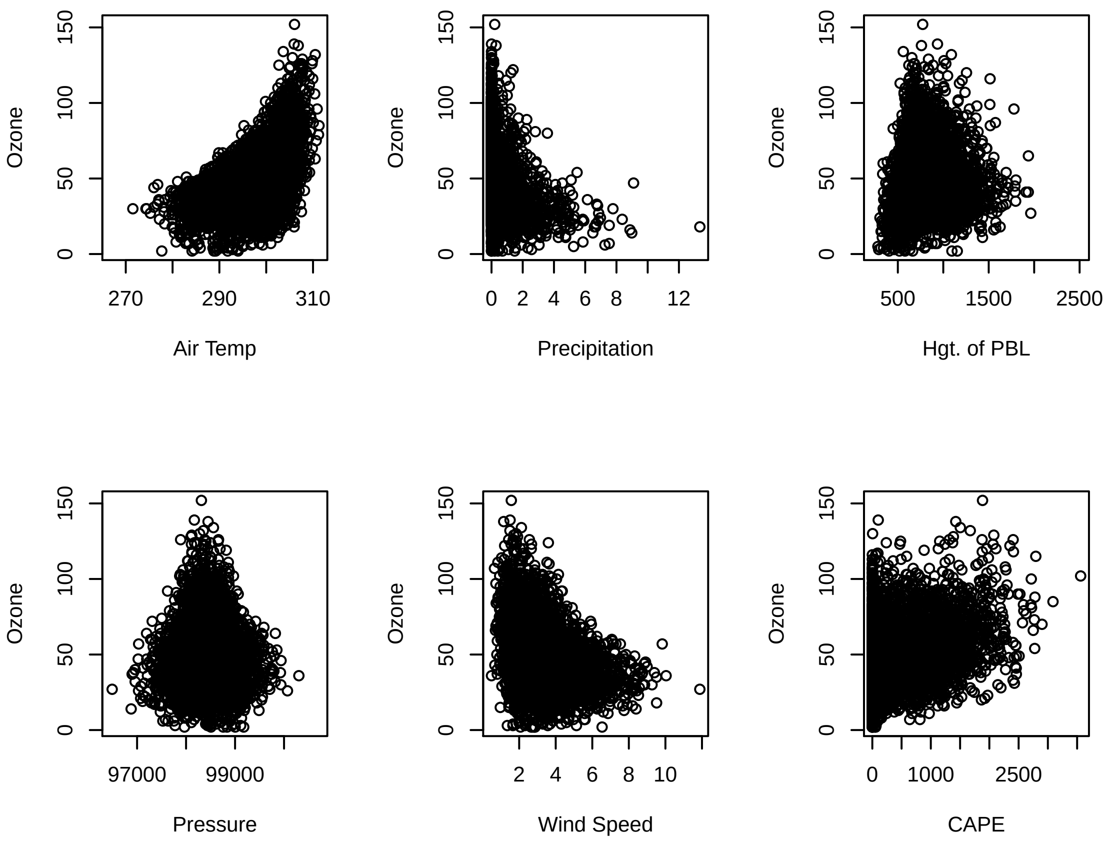
Simulation Experiment

5 covariates, a non-linear response in extreme behavior.

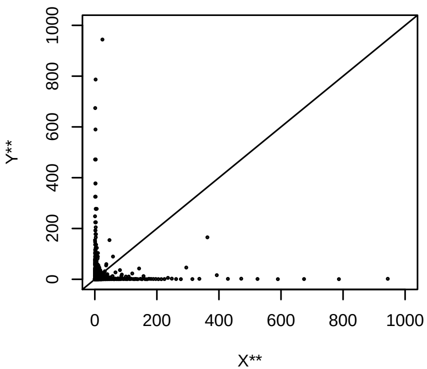


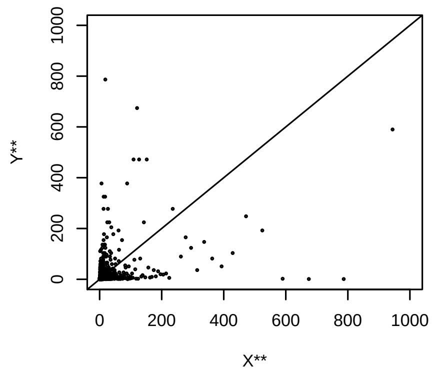
Case Study: Ozone for Atlanta

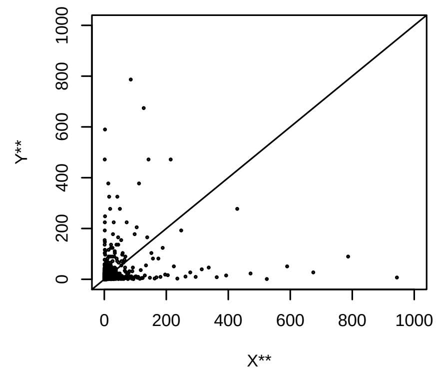
Covariates: met data from NARR reanalysis, approx 60.
Response: max of 8-hour moving average.



Preliminary Results for Atlanta







Score: 212.2
CV: 196.77
convective cloud cover
PBL height

Score: 119.78
CV: 114.7
air temp
windspeed

Score: 80.69
CV: 105.24
air temp†
air temp squared
wind speed†
3-day total cloud cover†
wind direction†
pressure
3-day PBL height
cloudcover×windspeed†
precip indicator†
precip×total cloud cover

Project Status

- Short/Mid-term
  - Perform large-scale data mining, consider all 2-way and 3-way interactions.
  - Perform a directed search based on input from Colette, Alma.
  - Repeat analysis for Charlotte, test robustness of results.
  - Repeat for PM_{10}.
  - Communicate findings to modelers.
- Longer term
  - Perform a spatial analysis.
  - Apply to model output and compare results to observations.