Stat547L: Spatio–temporal methods in environmental epidemiology: Lecture 23

Gavin Shaddick[1] & Jim Zidek [2]

[1]University of Bath
[2]University of British Columbia

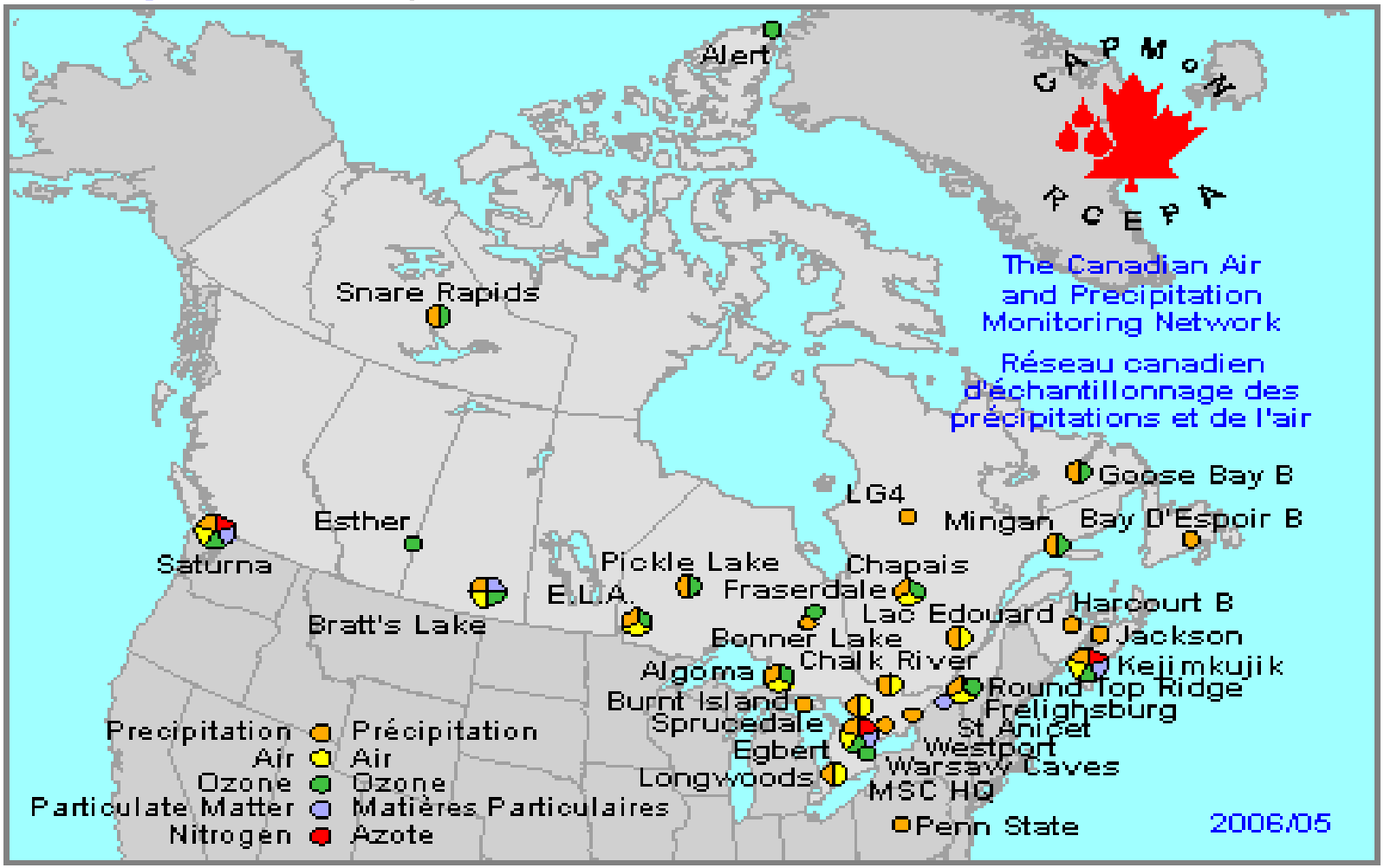
Term 2, 2012/13

SPECIAL TOPIC:

Designing monitoring networks

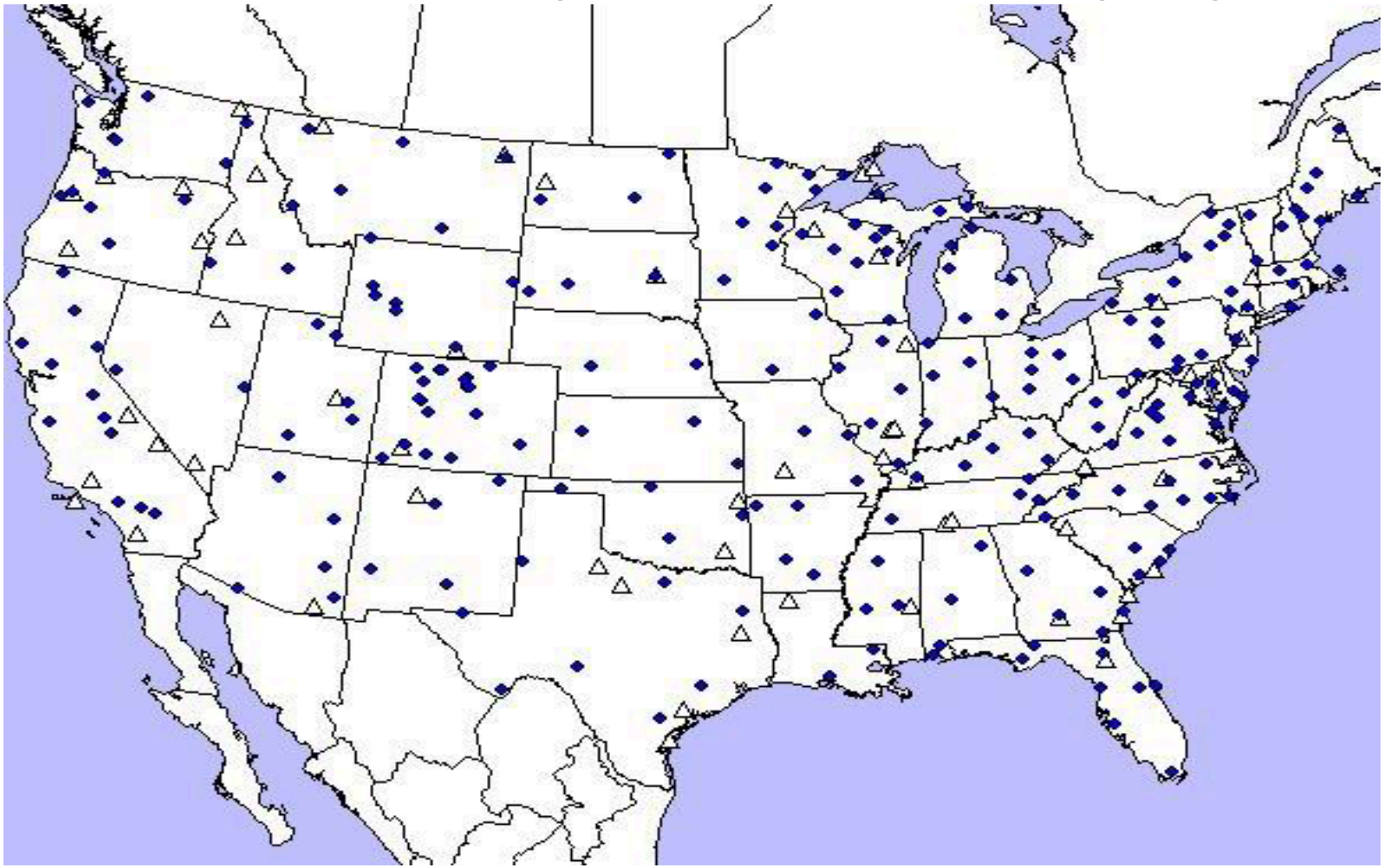
Designing networks

Example: the Capmon network.



The NADP/NTN network

Monitors multivariate responses related to “acid precipitation”



NOTES on Capmon:

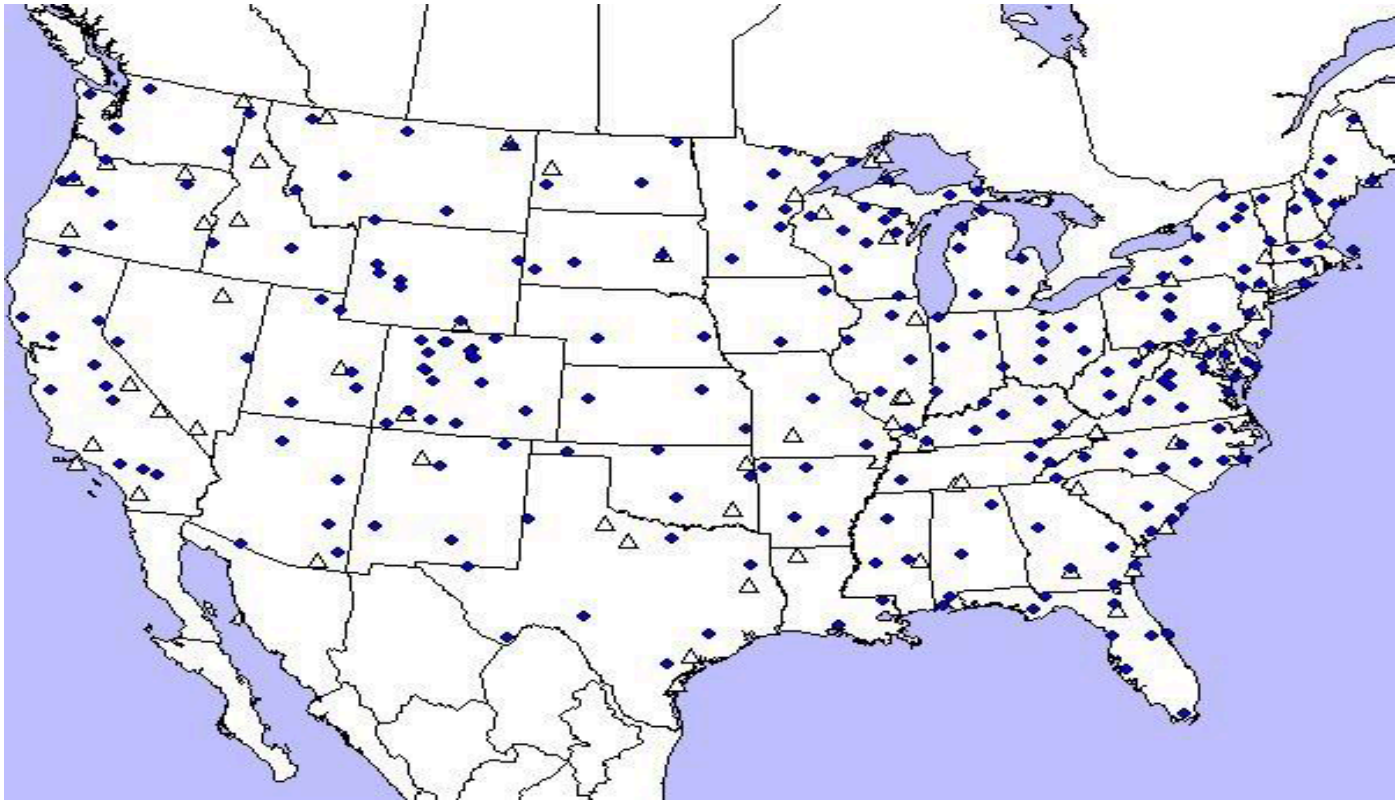
- No sense an “optimal” network for monitoring the environment.
- For administrative simplicity Capmon was a merger of three networks, each setup to monitor acid precipitation when that topic was fashionable.
- For simplicity, the sites were then adopted for other things, e.g, air pollution

---

Lessons learned: A network’s purposes often diverse and unforeseen.

The NADP/NTN network

Monitors multivariate responses related to “acid precipitation”– another merger–better defined siting rules!



Rules governing siting and types of NDP/NTN monitors:

"1. The COLLECTOR should be installed over undisturbed land on its standard 1 meter high aluminum base. Naturally vegetated, level areas are preferred, but grassed areas and up or down slopes up to 15% will be tolerated. Sudden changes in slope within 30 meters of the collector should also be avoided. Ground cover should surround the collector for a distance of approximately 30 meters. In farm areas a vegetated buffer strip must surround the collector for at least 30 meters." ⋮

Why monitor?

General objectives: To:

- measure process responses at critical points:
  - Near a new smelter using arsenic
- enable predictions of unmeasured responses
- enable forecasts of future responses
- provide process parameter estimates
  - physical model parameters
  - stochastic model parameters eg. covariance parameters
- address societal concerns

Specific objectives: To:

- detect non-compliance with regulatory standards
- enable health effect assessments to be made
  - & provide good estimates of relative risk
  - determine how well sensitive sub-populations are protected
  - can include all life, not just human
- to assess temporal trends
  - are things getting worse?
  - is climate changing?

Overall purposes

- to explore/reduce uncertain uncertain aspects of the environmental processes
  - one form of uncertainty (***aleatory***) cannot be reduced (outcome of fair die toss
  - the other (***epistemic***) (whether the die is fair) can increase or decrease. Implication: even an optimum design must be regularly revisited

What's "uncertainty"?

- Laplace: "Probability is the language of uncertainty"
- DeFinetti: "In life uncertainty is everything"
- Statisticians: "variance" or "standard error"
- Kolmorov & Renyi: "Entropy"

Exercises

23.1 Suppose $X \sim N(0, 1)$. Prove that uncertainty about X, i.e. Var(X——X—¡C) is increasing as a function of C.

23.2 Suppose $X \sim N(\eta, 1)$. Prove that uncertainty about X, i.e. $Var(X \mid\mid X \mid< C)$ is increasing as a function of C. Warning:Very hard!

Was for many years and until very recently an unsolved problem posed by van Eeden and Zidek. Rewards were offered and increased until there was winner in 2009.

A large check for $100 was presented at an official ceremony at UBC that year.


SURPRISE ANNOUNCEMENT !!!
He will now be receiv check and explain how
he got his answer.
Jiahua Chen
$100.000
James V. Zidek
Inbox - Thunderbird

Possible design criteria

"Gauge" (add monitors to) sites that

- that maximally reduce uncertainty at their space-time points
  - measuring their responses eliminates their uncertainty
- best minimize uncertainty about their cousin's responses
- best inform about process parameters
- best detect **non-compliers**

Designer challenges:

- multiplicity of valid objectives
- unforeseen & changing objectives
- multiple responses at each site: which to monitor?
- must include prior knowledge & prior uncertainty
- should use realistic process models. (How?)
- must integrate with existing networks
- must deal with reality!!!

Exercises

23.3 How might design criteria be arrived at in practice? Who should be responsible for setting them?

23.4 Monitor placement should recognize such things as the geographical distribution of impacted populations (eg trees or fish). How can an optimal design be determined in such a context? Research question!

23.5 Develop a design theory in a non-Gaussian context. Research question!

Approaches to design

Space-filling designs

Probability based designs

- simple random sampling
- stratified, multistage designs
- e.g. (1) EPA's survey of lakes; (2) the EMAP project

Model Based

- Regression model approach
  - eg to estimate the slope put 1/2 the data at each end of the data range
- Random fields (prediction, e.g. entropy) approach

Other. In particular Zhengyuan Zhu (UNC) incorporates both of the latter, prediction and parameter estimation.

Entropy based approach

"Gauges" sites with greatest "uncertainty"

- **uncertainty = entropy**
- **maximally reduces uncertainty about "ungauged" sites**
- **best estimates predictive posterior distribution under entropy utility**

By - passes specification of objectives

Long history[1], currently popular

[1] General: Good [1952], Lindley [1956], Shewry and Wynn [1987]. Network design: Caselton and Zidek [1984],Sebastiani and Wynn [2002],Zidek et al. [2000]

What's entropy?

Let $p = P(E)$= probability an uncertain event E occurs (heads on possibly bent coin). That uncertainty is reduced to none when outcome known, a reduction of say (for some function ϕ)
$\phi(p)$ if E occurs
$\phi(1-p)$ if not.

The expected reduction in uncertainty is
$p\phi(p) + (1-p)\phi(1-p)$
Simple assumptions imply:

$$\boxed{\phi(p) = log(p)}$$

Thus entropy reduction due to knowledge of E's occurrence (= "**uncertainty**" about E) is the **entropy** for the two point distribution $(p, 1-p)$:

$$p \log(p) + (1-p) \log(1-p)$$

Relative entropy

How much is that entropy?
Needs a reference level. Complete uncertainty about the coin, how its to be tossed and so on would point to a two point distribution $(q, 1-q)$ with $q = 1/2$. Thus the relative entropy would be
$I(p, q) = p \log(p/q) + (1-p) \log\{(1-p)/(1-q)\}$ Kullback-Leibler's measure of deviation of $(p, 1-p)$ from its reference level (that corresponds to a "state of equilibrium" in physics (thermodynamics).

Multiple events

$$I(p, q) = \sum_i p_i \log \{p_i / q_i\}$$

Continuous variables

Start with $p_i \sim f(x_i)dx_i$ & $q_i \sim g(x_i)dx_i$ as approximations. Then as $dx_i \to 0$, this entropy converges to

$$I(f,g) = \int f \log f/g$$

.

Commonly $g \equiv 1(\textit{unitsoff})$. In any event, f/g is a unitless quantity. Moreover Jacobean cancels under transformations of x making entropy an "intrinsic" measure of uncertainty – not scale dependent.

Using entropy

Adopt a Bayesian framework. **Let:**

- $Y =$ process response vector at future time T+1 including all sites (gauged & ungauged)
- $D =$ set of all available data upon which to condition and get posterior distributions
- h_1 & h_2 be baseline reference densities against which to measure uncertainty.
- Finally:

$$\begin{aligned} H(Y \mid \theta) &= E[-\log(f(Y \mid \theta, D)/h_1(Y) \mid D] \\ H(\theta) &= E[-\log(f(\theta \mid D)/h_2(\tilde{\theta})) \mid D] \end{aligned}$$

Then we get fundamental identity (**Exercise**):

$$H(Y, \theta) = H(Y \mid \theta) + H(\theta)$$

Design goal

Can add or subtract sites in general. Let's focus on

adding new sites to an existing network

- $Y = (Y^{(1)}, Y^{(2)}) =$ all site responses, time T +1
- $Y^{(2)}$ for site currently gauged (time T)
- $Y^{(1)}$ for sites currently ungauged (time T)

DESIGN GOAL: Partition $Y^{(1)} = (Y^{(rem)}, Y^{(add)})$ at time T so that

- $Y^{(rem)}$: future ungauged sites
- $Y^{(add)}$ future new network stations.

Entropy decomposition thm

Let U=$Y^{(rem)}$; $G = (Y^{(add)}, Y^{(2)})$; Y=[U,G]

Fundamental identity:

$$\boxed{\text{TOT} = \text{PRED} + \text{MODEL} + \text{MEAS}}$$

where

$$\begin{aligned} PRED &= E[-\log(f(U \mid G, \theta, D)/h_{11}(U)) \mid D], \\ MODEL &= E[-\log(f(\theta \mid G, D)/h_2(\theta)) \mid D], \end{aligned}$$

and

$$MEAS = E[-\log(f(G \mid D)/h_{12}(G)) \mid D].$$

$$\boxed{\text{Theorem: Maximizing MEAS=Minimizing MODEL + PRED}}$$

Response distribution

After removing regular temporal components and transforming responses as necessary assume the enviro.stat model:

$$X \mid \beta, \mathbf{\Sigma} \sim N(Z\beta, I_n \otimes \mathbf{\Sigma})$$

$$\beta \mid \mathbf{\Sigma}, \beta_0, F \sim N(\beta_0, F^{-1} \otimes \mathbf{\Sigma})$$

$$\mathbf{\Sigma} \sim GIW(\Psi, \delta) \text{ \# Generalized Inverted Wishart distribution}$$

Generalized IW properties:

Begin with notation:

$$\Sigma = \begin{pmatrix} \Sigma^{[u]} & \Sigma^{[ug]} \\ \Sigma^{[gu]} & \Sigma^{[g]} \end{pmatrix}$$

invoke the Bartlett decomposition: $\Sigma = T \Delta T'$ where

$$T = \begin{pmatrix} I & \Sigma^{[ug]}(\Sigma^{[g]})^{-1} \\ 0 & I \end{pmatrix}$$

$$\Delta = \begin{pmatrix} \Sigma^{[u]} - \Sigma^{[ug]}(\Sigma^{[g]})^{-1}\Sigma^{[gu]} & 0 \\ 0 & \Sigma^{[g]} \end{pmatrix}$$

Now let $\Gamma^{[u]} = \mathbf{\Sigma}^{[u|g]} = \mathbf{\Sigma}^{[u]} - \mathbf{\Sigma}^{[ug]}(\mathbf{\Sigma}^{[g]})^{-1}\mathbf{\Sigma}^{[gu]}$; $\tau^{[u]} = (\mathbf{\Sigma}^{[g]})^{-1}\mathbf{\Sigma}^{[gu]}$.

Then $\mathbf{\Sigma} \sim GIW(\Psi, \delta)$ means for appropriate hyperparameters:

$$\mathbf{\Sigma}^{[g]} \sim GIW(\Psi^{[g]}, \delta^{[g]})$$

$$\Gamma^{[u]} \sim IW(\Lambda_0 \otimes \Omega, \delta_0)$$

$$\tau^{[u]} \mid \Gamma^{[u]} \sim N\left(\tau_0 u, H_0 \otimes \Gamma^{[u]}\right)$$

Predictive distribution becomes

$$\begin{aligned}\left(Y^{[u]} \mid D, \mathcal{H}\right) \quad \sim \quad & \left(Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \times \\ & \prod_{j=1}^{k-1} \left(Y^{[g_j^m]} \mid Y^{[g_{j+1}^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \\ & \times \left(Y^{[g_k^m]} \mid D, \mathcal{H}\right).\end{aligned}$$

$$\left(Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \sim$$

$$t_{n\times up}\left(\mu^{[u|g]}, \text{Dispersion}, \delta_0 - up + 1\right).$$

$$\text{Dispersion} = (\delta_0 - up + 1)^{-1}\Phi^{[u|g]} \otimes (\Lambda_0 \otimes \Omega)$$

Thus the conditional entropy for $Y^{[u]}$ that has to be partitioned into 'add' and 'rem' sites is

$$H\left[Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D\right] = \frac{p}{2}\log|\Lambda_0| + \text{irrelevant terms}$$

Thus recalling the Bartlett decomposition, we can optimize the choice of the 'add' sites at time T+1 by maximizing $|\Lambda_0[add, add]|$, the sub–determinant of $|\ \Lambda_0\ |$ corresponding to the 'add' sites in the partitioned $Y^{[u]}$.
That will simultaneously minimize the entropy left in the 'rem' sites.

‘Add’ computation

- **NP-Hard:** No exact algorithms for big networks
- Inexact Methods:
  - Greedy
  - Greedy + Swap
- Exact Methods:
  - Complete enumeration
  - Branch and bound

How many sites?

Compute entropy/number of sites as the number of sites varies.
Eventually this reaches a max (bang for the buck)and then declines.
Indicates when to stop on redesign.

Example:

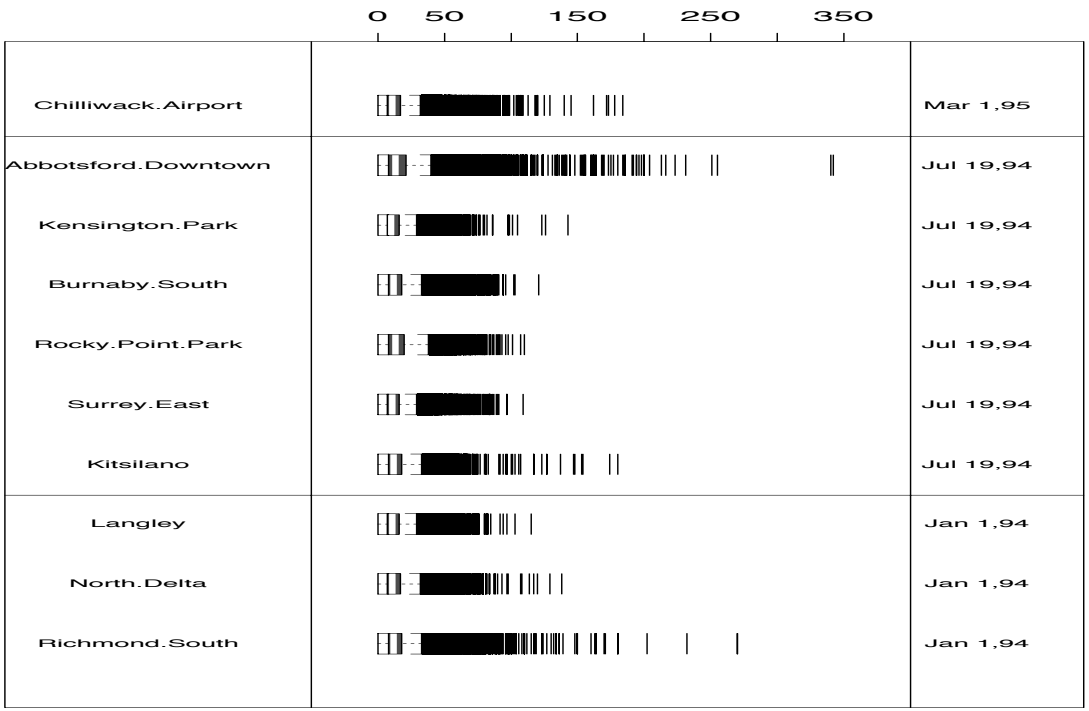
Hypothetical redesign of Vancouver's hourly PM_{10} field. Existing 10 monitoring sites are to be increased to 16 by selecting 6 new stations from among 20 possible sites. Use the entropy approach and

Normal- generalized inverted Wishart predictive distribution

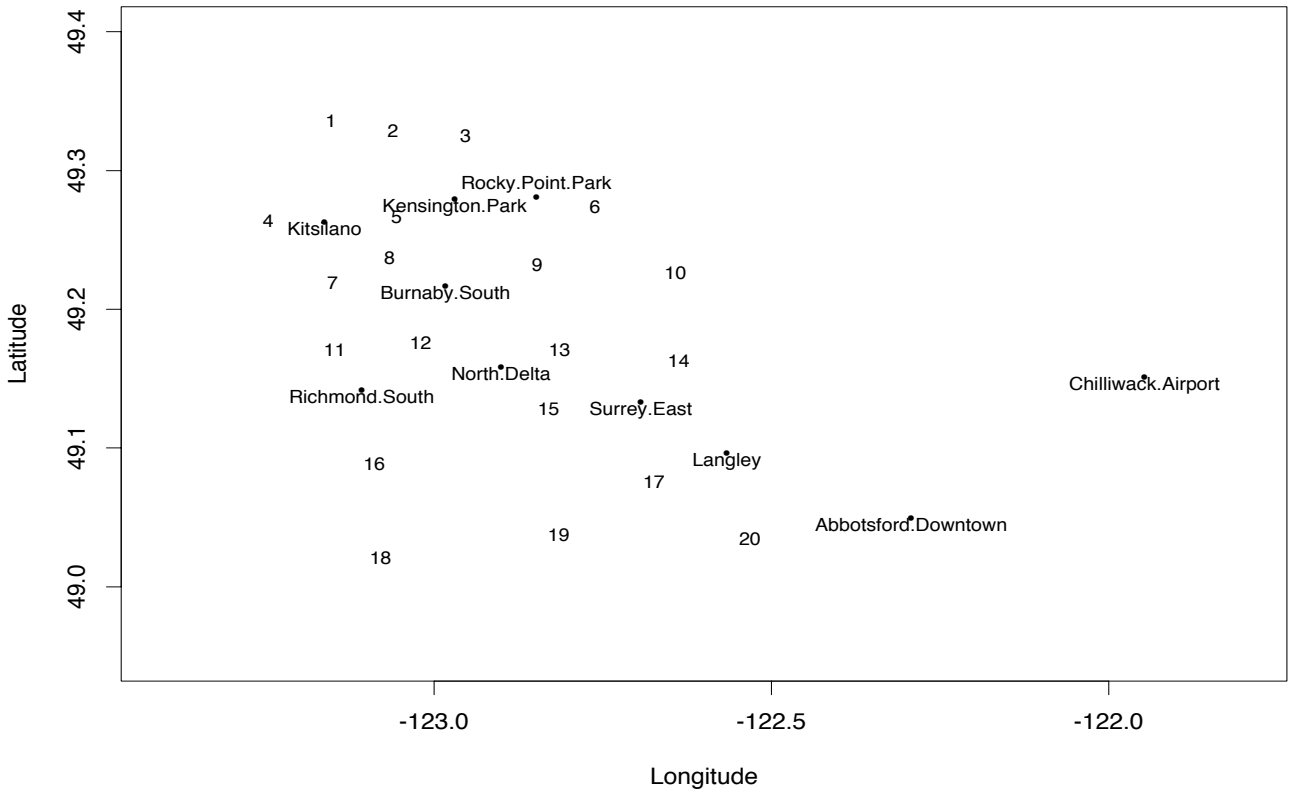
- **needs "whitened" residuals; space - time interaction $\rightarrow$ use of 24 (hour) dimensional multivariate AR(1) model**
- **different 10 - station startups $\rightarrow$ monotone ("staircase") data structure $\rightarrow$ generalized inverted Wishart distribution $\rightarrow$ different d.f. for each staircase step**
- **select the 6 new stations with jointly maximum conditional entropy**

Results:

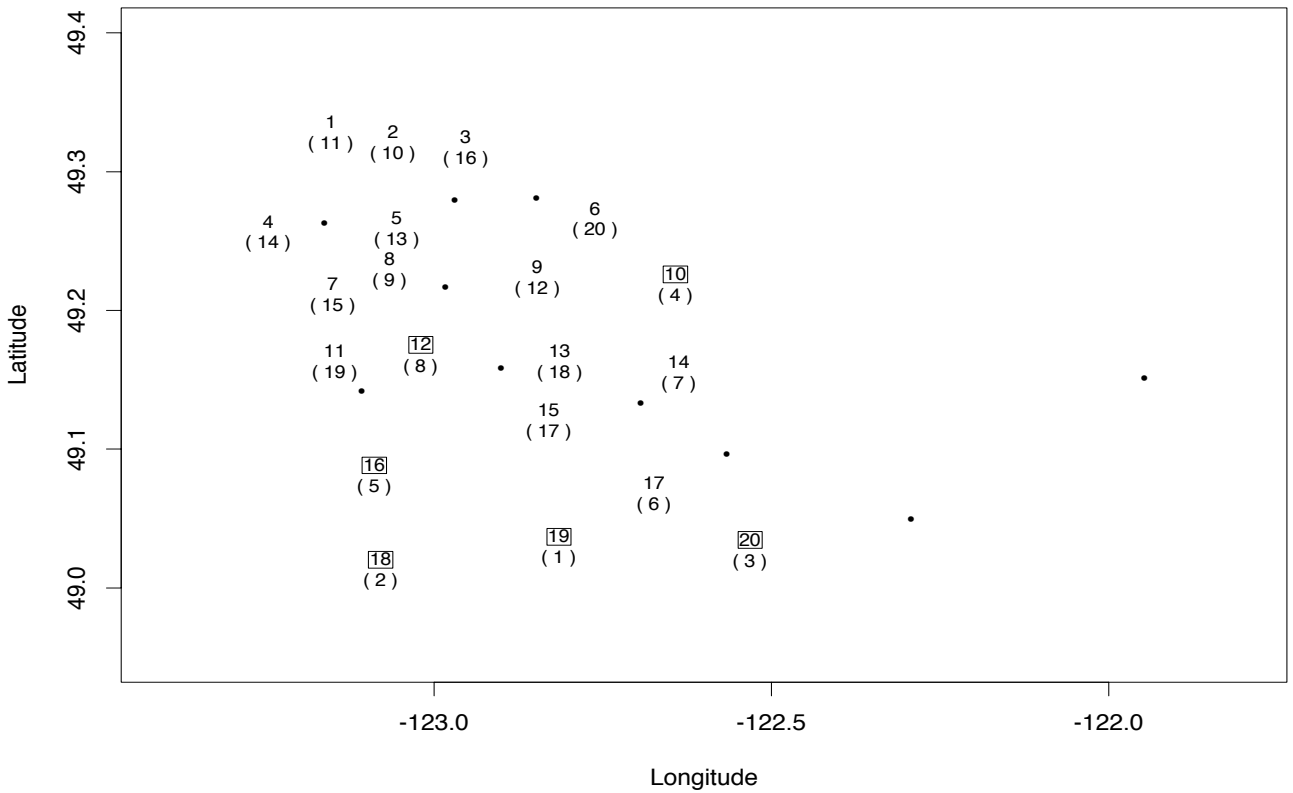
Here is a look at PM_{10}'s levels at the 10 existing stations. Note differing startup times.



The 10 PM_{10} monitoring sites (the ones with names) & prospective new locations.



Locations of the old and newly selected ‘add’ sites (square brackets). The ranks of the 20 sites by estimated variance is in curved brackets.



Non-Compliance detection designs

Probability, hence best design, day - dependent!

- **which day?**
- **a simulated future day? Average day? Bad day?**

How implemented?

- **monitor sites most likely to comply?**
- **do not monitor sites least likely?**
- **what about existing sites?**

Entropy vs noncompliance

Example: how well would Vancouver's 6 site, entropy-based, addition compare to an optimal noncompliance based addition?

Use 10 station hourly data for PM_{10}, February 28, 1999 & hierarchical Bayes predictive distribution

CRITERION:

$$\textbf{argmax PR}\{\textbf{daily max } PM_{10}\ \mathbf{Y}^{6added} \geq \mathbf{50}\ (\mu g\ m^{-3})\}$$

NOTES: Entropy does not work nearly as well on August 1, 1998!

Results for Vancouver redesign for noncompliance

The selected new sites are now determined pretty much by their posterior estimated variances. That is because the spatial correlation is now quite weak.

Special problem: monitoring extreme values

Regulatory criteria metrics (risk) usually involves extremes. **Example:** EPA'S PM_{10} criterion:

For particles of diameters of 10 micrometers or less:

Annual Arithmetic Mean: 50 μg m^{-3}

24 - hour Average: 150^{FN} μg m^{-3}

The three year average of 98-th annual percentiles of 24 hour averages must be ≤ 150 ($\mu g\ m^{-3}$) at all sites in an urban area. Complex metric $\Rightarrow$ need predictive distribute to simulate its distribution!

Bad news in brief re extreme value situation

1. **Insufficient data, spatial and temporal.**
2. **Extremes have small inter - site dependence**
   - **between some site pairs, not others**
3. **Conventional approaches fail**
4. **Multivariate extreme value distributions - not tractable**
   - **conditional computation (e.g. entropy) difficult**
   - **simulating extreme fields hard**
5. **Elusive design objective**

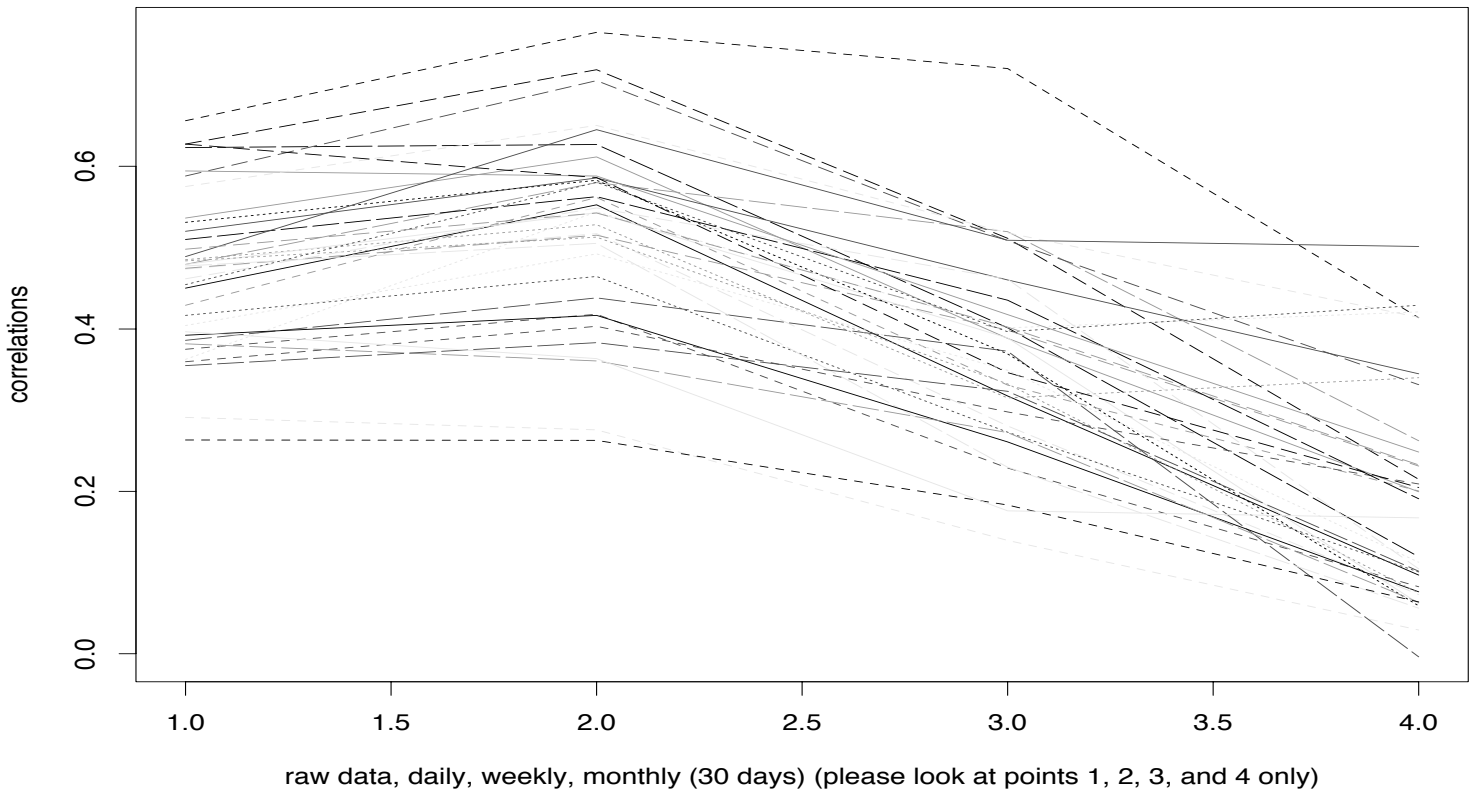
Good news in brief

Joint distribution of extremes approximately a log multivariate t distribution. Hence can:

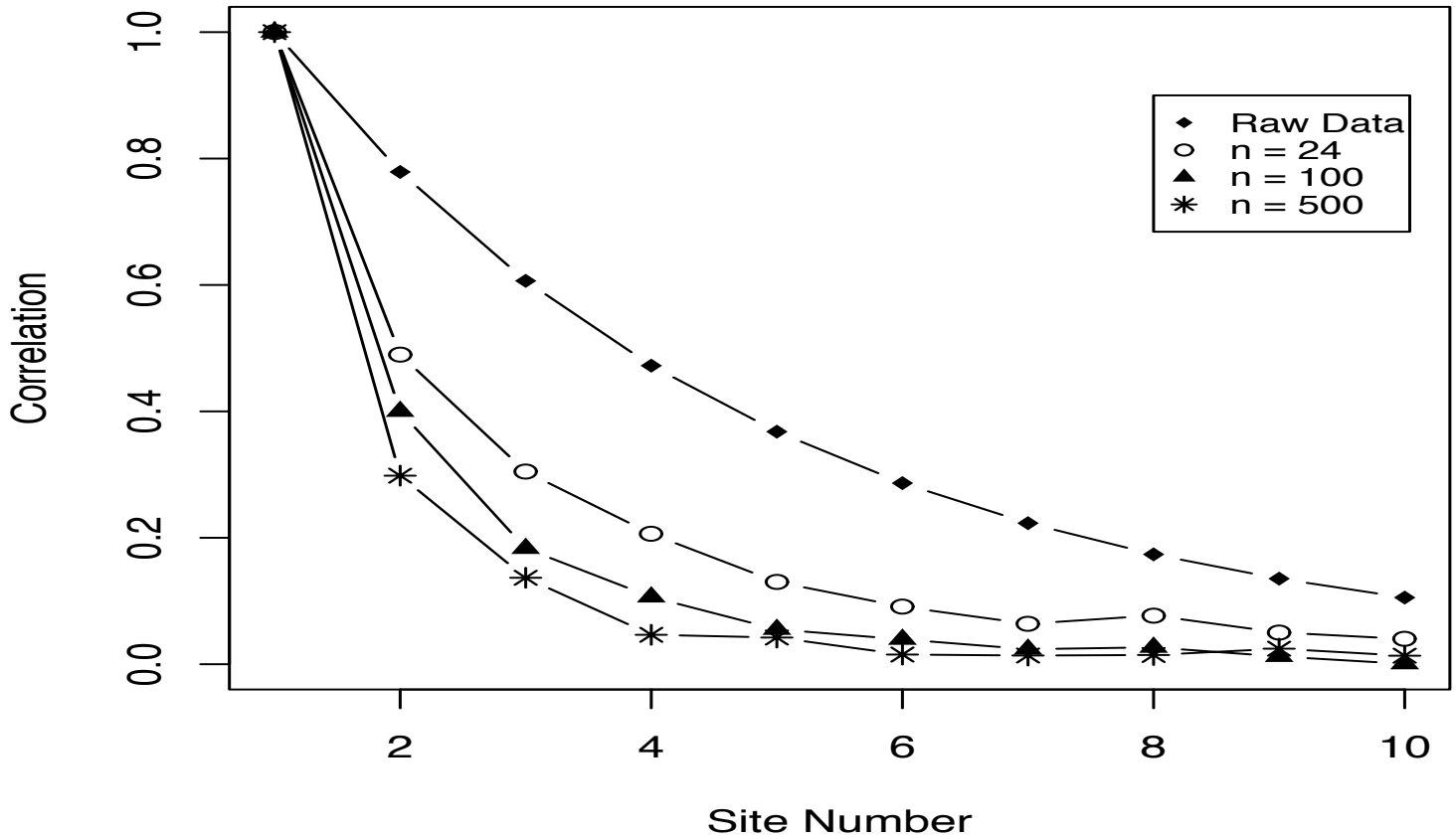
1. **have convenient conditional, marginal distributions**
2. **accommodate existing sites and historical data**
3. **permit simulation of complex metric distributions**
4. **have explicitly computable entropy's, regression models, etc**
5. **can enable "elusive objectives issue" to be bypassed**

More detail: Small inter-site correlations

Inter- site dependence declines with increases in extreme's"range" for many, not all site pairs [London and Vancouver analyses]. Figure shows Vancouver's for PM_{10} decline with max range.



***Simulation study*[2] :** *multinormal responses; maxima with varying ranges at 10 sites. Multivariate t results show smaller loss of dependence. Inter-site correlations for maxima for simulated fields of extremes. Big n = light tails.*

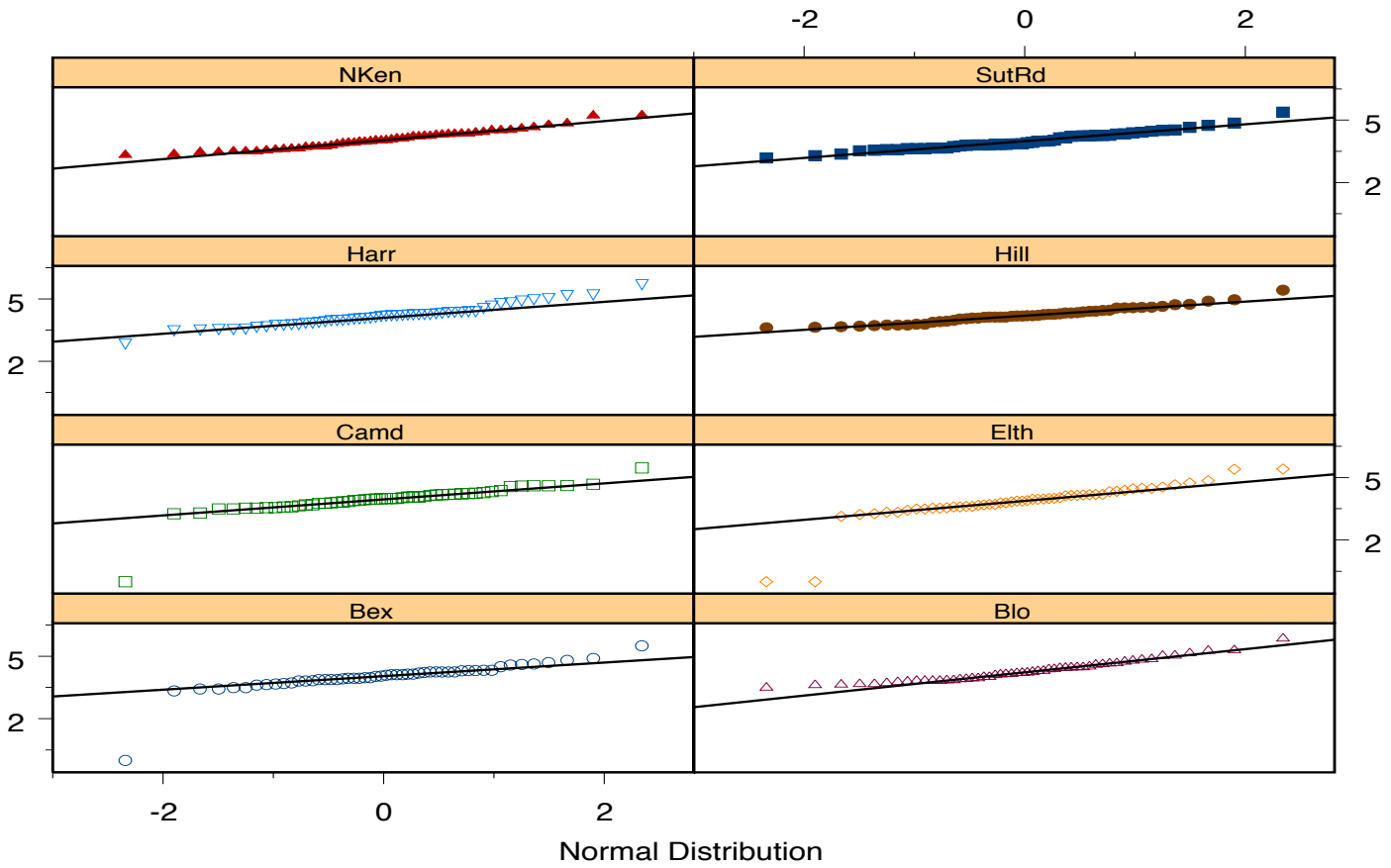


Exercises

23.6 Try the simulation experiment yourself and confirm that for heavier tails lead to increased intersite correlations

Approach to monitoring extremes

Empirical results $\mapsto$ log multivariate - t distribution as approximation to joint distribution of extremes field. QQplots for weekly maxima of hourly log PM_{10} London 1997 data $\rightarrow$ **marginal normality of extremes**:



T approx approach cont'd

Empirical results $\rightarrow$ well-calibrated 95% (etc) prediction intervals. Supports use of multivariate approximation.

Credibility Level	Mean	Median
30%	35	35
95%	96	97
99.9%	99.9	1

Table: **Summary of coverage probabilities at different credibility levels for the simulated precipitation data over 319 grid cells, Canadian Climate Model**

The good news

Use of log multivariate t distribution for extreme fields promising. But:

- **how far can approximation go $\rightarrow$ need for theory**
- **test approximation case-by-case**
- **no substitute for knowledge of latent processes**
- **need to compare regular - and extreme-entropy designs.**

Recommendations to regulator

Spend some data dollars assessing current designs

Keep things simple! Select metrics susceptible to analysis.

Think about/articulate design purposes

Conclusions

Current urban networks may not be dense enough for adequate surveillance

Conventional designs inadequate but MaxEnt may be adapted/used

More knowledge of latent processes needed

More attention to design criteria for extremes needed

Designing for extremes $\rightarrow$ significant challenges

William F Caselton and James V Zidek. Optimal monitoring network designs. *Statistics & Probability Letters*, 2(4):223–227, 1984.

Howard Chang, Audrey Qiuyan Fu, Nhu D Le, and James V Zidek. Designing environmental monitoring networks to measure extremes. *Environmental and Ecological Statistics*, 14(3):301–321, 2007.

Irving John Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

Paola Sebastiani and Henry P Wynn. Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1): 145–157, 2002.

Michael C Shewry and Henry P Wynn. Maximum entropy sampling. *Journal of Applied Statistics*, 14(2):165–170, 1987.

James V Zidek, Weimin Sun, and Nhu D Le. Designing and integrating composite networks for monitoring multivariate gaussian pollution

fields. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(1):63–79, 2000.