It's not all about big data, but some of it is[1]

Thomas A. Louis, PhD

Department of Biostatistics
Johns Hopkins Bloomberg SPH
tlouis@jhu.edu

Expert Statistical Consultant
Center for Drug Evaluation & Research
U.S. Food & Drug Administration
Thomas.Louis@fda.hhs.gov

[1] Presented during the joint Conference of the Sub-Saharan Africa Network of the International Biometrics Society, and DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics

Principal Points

- Big data are (is) everywhere
- They (it) present opportunities and challenges
- My focus is on challenges generated by missing information on
  - The sampling plan
  - The reference population
- Theory and examples highlight the challenges and confer a degree of hope
- Guidance and a reprise provide the capstone

The drumbeat Big Data ::: Found data ::: Data Exhaust ::: All data

- Popular media and science publications sound the drum,

  > 'Big Data' will drive our future, from translating genomic information into new cancer therapies to harnessing the Web for untangling complex social interactions or detecting infectious disease outbreaks[2]

- 'Datafication' of everything
- Features of 'pure' big data
  - VVV: Volume, Velocity, Variety
  - Organic creation
  - Passive data collection
  - Instability
- The Statistico-centric world must cohabitate with the data-centric world

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete
https://www.wired.com/2008/06/pb-theory/

[2] Davidian & Louis (2012). Why Statistics? *Science*, 336: p12)

4 Big Data to the rescue?

NEWS ANALYSIS

Apple and IBM say big data will save lives

Apple will let us know more on its plans for health on June 8 when WWDC 2015 begins. Credit: Apple

An Apple (and IBM) each day may keep you healthy

Computerworld | Apr 14, 2015 6:00 AM PT

4 Big Data to the rescue?

NEWS ANALYSIS

Apple and IBM say big data will save lives

Apple will let us know more on its plans for health on June 8 when WWDC 2015 begins. Credit: Apple

An Apple (and IBM) each day may keep you healthy

Computerworld | Apr 14, 2015 6:00 AM PT

Take a daily, ‘big data’

Care is needed

- 'Big data' does not imply big, relevant or valid information
- Science requires uncovering causal relations; while Big Data has produced interesting and important predictions & associations, care is needed to move from these to explanation, causation and transportability[3]
- Issues and challenges include,
    - Instability of the data generating process
    - Bias, confounding and poorly informed representation as threats to validity
- Modern techniques can improve validity, but are unlikely to be fully successful
- There are definite roles for Big Data, but in many contexts they should supplement/complement and not replace well-curated data

[3] Pearl J, Bareinboim E (2014). External Validity: From do-calculus to Transportability across Populations. *Statistical Science*, 29: 579–595

6 Importance of the sampling plan

- The sampling plan determines the scope of and methods for inference
- There is always a sampling plan, and here are some examples:
  - Random, stratified random, cluster, sno-ball
  - Haphazard, convenience, as they arrive (a series)
  - "I have no idea"
- Selection effects, informative dropouts and other types of missing data affect sample representation
- If you know the sampling weights, even for the observed sample, you have a representative sample, of some population
  - Need an identified reference population to complete the connection

Importance of the sampling plan

- The sampling plan determines the scope of and methods for inference
- There is always a sampling plan, and here are some examples:
  - Random, stratified random, cluster, sno-ball
  - Haphazard, convenience, as they arrive (a series)
  - "I have no idea"
- Selection effects, informative dropouts and other types of missing data affect sample representation
- If you know the sampling weights, even for the observed sample, you have a representative sample, of some population
  - Need an identified reference population to complete the connection

My focus is on missing or incomplete information on the sampling plan and reference population

Estimating a Population Mean

(Imagined Hospital Length of Stay, LOS, data)

- Estimate the average LOS for hospitals in a specific domain
- Assume the target population consists of 5 hospitals, that a random sample of n_j medical records from hospital j is obtained
  - $\sigma_j^2 \equiv \sigma^2$

	Observed Information				**Population Information**		
Hospital	# sampled n_j	% of total sample $100f_j$	Mean LOS (Y_j)	Sampling Variance	Hospital size	% of total pop. $100p_j$	Patient relative propensity (f_j/p_j)
1	30	**20**	25	$\sigma^2/30$	100	**10**	**2.00** $= 20/10$
2	60	40	35	$\sigma^2/60$	150	15	2.67
3	15	**10**	15	$\sigma^2/15$	200	**20**	**0.50** $= 10/20$
4	30	20	40	$\sigma^2/30$	250	25	0.80
5	15	10	10	$\sigma^2/15$	300	30	0.33
TOTAL	150	100			1000	100	

- The sample is not self-weighting; some patient relative propensities $\neq$ 1.00
- It is representative because the relative propensities are known

Weights, weighted averages and relative variances

Estimator	Hospital-specific Weights ($\mathbf{w}$)	$\hat{\mu}(\mathbf{w})$	Variance Ratio $100\times$(Var/minVar)
Minimum Variance	.20 .40 .10 .20 .10	29.5	100
Equally weighted	.20 .20 .20 .20 .20	25.0	130
Unbiased	.10 .15 .20 .25 .30	23.8	172

- 'Minimum variance' and 'Equally Weighted' are available from the sample information
- 'Unbiased' depends on the relative propensities, which require frame and sampling plan information

Are non-probability samples informative?

- Many state that nonprobability, 'volunteer samples,' can't be used for population estimates because the necessary weights aren't available,

  > The debate over probability vs. nonprobability samples is about representation.[4]

- However, would you rather have 60% response rate from a well-designed and conducted (Gallup) survey or a 95% rate from a self-selected group?
  - **Advantage Gallup:** The 60% is also self-selected, but information on the relation of respondents to non-respondents is available from the sampling frame and generalizing from the sample is possible
  - **Non-probability has potential:** There may be other data that can be used to develop reasonable weights for some reference population
    - Use all data (big, small, in-between) to help identify the population and compute weights
- Analogously, in clinical trials most causal questions are not protected by randomization, are not ITT, but careful, causal analysis can be valid
  - For analogies between non-probability surveys and causal inference, see[5]

[4] Keeter (2014). Change Is Afoot in the World of Election Polling *amstat news,* October: 3-4.

[5] Mercer, Kreuter, Keeter, Stuart (2017). Theory and Practice on nonprobability surveys, Parallels between causal inference an survey inference (with discussion). *Public Opinion Quarterly*, 81: 250–279.

Xiao-Li Meng's Cautionary Tale[6,7]

(A big sample size, n, may not save the day)

- Compare the MSE for two estimators of the finite population mean ($\bar{Y}_N$), N large

  $\bar{y}_{srs}$: Sample mean of a simple random sample of size $n_{srs} = 100$
  $\bar{y}_{sel}$: A self-selected, web sample of size n_{sel}

- With $\rho(\mathbf{Y}, \boldsymbol{\pi}) = \text{cor}(\mathbf{Y}, \text{inclusion propensity}) = 0.05$, and $frac = n_{sel}/N$,

$$\text{MSE}_{sel} \leq \text{MSE}_{srs} \iff frac \geq 20\%$$

- For example, $N = 50M$ requires $n_{sel} \geq 10M$ to beat the SRS with $n_{srs} = 100$ (!)
- Good information on $\rho(\mathbf{Y}, \boldsymbol{\pi})$ is needed to rescue the situation

A large sampling fraction, n/N, may not be protective

- More on this later

[6] Meng's discussion of Keiding&Louis (2016)

[7] Meng (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 Presidential Election. *Annals of Applied Statistics*, 12: 685–726.

Validation from population-level databases

A finding that did not generalize

- In the Nordic countries individual record linkage to detailed population registries sometimes allows validation of the representativity of a study cohort, which is always at least partly based on volunteers

- Andersen et al. (1998)[8] compared mortality of participants in 3 cohorts recruited in the Copenhagen area to the general mortality in that area
- There is a risk of bias if other causes for the disease under study or confounders are not taken into account and are differently distributed among the participants and the target population
- Many factors associated with disease and death differ between participants and non-participants either because they are implicit in the selection criteria or because of the self-selection
- The analysis showed survivor selection in all cohorts (recruited participants being healthier at baseline than non-recruited individuals), which persisted beyond ten years of observation for most combinations of age and sex

[8] (1998) A comparison of mortality rates in three prospective studies from Copenhagen with mortality rates in the central part of the city, and the entire country. *European J. of Epidemiology*, 14: 579–585

Validation from population-level databases

A finding that did generalize

- Results from clinical trials on breast-conserving operations appear applicable to all Danish women[9]
- The Danish Breast Cancer Cooperative Group (DBCG) coordinates breast cancer therapy in Denmark, where almost all women are treated for free at the public hospitals
- Many RCTs on adjuvant therapy have been conducted with sampling frame all Danish women, suitably stratified (e.g., by age and/or menopausal status)
- From 1982 to 1989 a randomized trial compared breast conserving surgery to total mastectomy, and subsequently breast conserving therapy was offered as option to qualifying patients across Denmark
- The population-based registry of the DBCG allowed population-based follow-up 1989-98, finding that:

  > Women younger than 75 years and operated on according to the recommendations, had survival, loco-regional recurrences, distant metastases and benefit from adjuvant radiotherapy closely matching the results from the clinical trial

- See also[10]

---

[9] Ewertz et al. (2008) Breast conserving treatment in Denmark, 19891998. A nationwide population-based study of the Danish Breast Cancer Co-operative Group. *Acta Oncologica*, 47, 682–690.

[10] Hviid, Hansen, and Frisch, Melbye (2019). Measles, Mumps, Rubella Vaccination and Autism: A Nationwide Cohort Study. *Annals of Internal Medicine*, 10.7326/M18-210.

Validation from population-level databases

A finding that did generalize

- Results from clinical trials on breast-conserving operations appear applicable to all Danish women[9]
- The Danish Breast Cancer Cooperative Group (DBCG) coordinates breast cancer therapy in Denmark, where almost all women are treated for free at the public hospitals
- Many RCTs on adjuvant therapy have been conducted with sampling frame all Danish women, suitably stratified (e.g., by age and/or menopausal status)
- From 1982 to 1989 a randomized trial compared breast conserving surgery to total mastectomy, and subsequently breast conserving therapy was offered as option to qualifying patients across Denmark
- The population-based registry of the DBCG allowed population-based follow-up 1989-98, finding that:

  > Women younger than 75 years and operated on according to the recommendations, had survival, loco-regional recurrences, distant metastases and benefit from adjuvant radiotherapy closely matching the results from the clinical trial

- See also[10]

It helps to be in Scandinavia!

[9] Ewertz et al. (2008) Breast conserving treatment in Denmark, 19891998. A nationwide population-based study of the Danish Breast Cancer Co-operative Group. *Acta Oncologica*, 47, 682–690.

[10] Hviid, Hansen, and Frisch, Melbye (2019). Measles, Mumps, Rubella Vaccination and Autism: A Nationwide Cohort Study. *Annals of Internal Medicine*, 10.7326/M18-210.

Big Data and Data Synthesis[11,12]

- Basic scenario: Have an internally valid, small(ish) study, and stable but possibly reduced dimension, external information
  - e.g, the joint distribution of a subset of the small study variables
- Approach: Constrain the small study estimates to be compatible with the externally determined relations
  - Analogous to stabilizing interior estimates in a contingency table by 'benchmarking' to marginal distributions estimated from other data
  - And to using external prevalence data to allow a case-control study to estimate a relative risk
- Key issue: Are stochastic features of the external data sufficiently similar to the relevant components of the small data to reduce MSE
  - Resonates with external validity, representativity of a sample, transporting within-sample estimates to a reference population, . . .

---

[11] Chatterjee, et al. (2016). Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-level information from External Big Data Sources (with discussion). *JASA*, 111: 107–131.

[12] Louis, Keiding (2016). Discussion of, Chatterjee et al. 123–124.

Design-based: The basic setup

- Finite population: $U = \{1, 2, \ldots, N\}$
- Values of interest: $Y_k,\ k \in U$
  - The Y_k are a set of fixed, but unknown numbers, not necessarily from a probability distribution
- Draw a sample $S \in U$ with,
  - $\mathrm{pr}(\text{unit } k \in S) = \pi_k > 0$ (can depend on covariates)
  - $\mathrm{pr}(k, \ell \in S) = \pi_{k\ell}$
  - $\mathrm{pr}(k_1, \ldots, k_n \in S) = \pi_{k_1 k_2 \ldots k_n}$
- Goal: Estimate a function of the Y_k, any function, but here the population total or mean

$$\text{total: } T(\mathbf{Y}) = \sum_{k=1}^{N} Y_k \qquad \text{mean: } A(\mathbf{Y}) = \frac{T(\mathbf{Y})}{N}$$

The weighting game

- Sample membership indicators:

$$Z_k = \begin{cases} 1, k \in S \\ 0, k \notin S \end{cases}$$

$$E(Z_k) = \pi_k \qquad E(Z_k Z_\ell) = \pi_{k\ell}$$

- The Z_k are random variables; the Y_k are constants
- The **Horvitz-Thompson**, unbiased estimate of T and nearly unbiased of A:

$$\hat{T} = HT\left[Y_k\right] = \sum_{k \in S} \frac{Y_k}{\pi_k} = \sum_{k \in U} \frac{Z_k Y_k}{\pi_k}$$

$$\hat{A} = \frac{\sum_{k \in S} \frac{Y_k}{\pi_k}}{\sum_{k \in S} \frac{1}{\pi_k}} = \frac{\sum_{k \in U} \frac{Z_k Y_k}{\pi_k}}{\sum_{k \in U} \frac{Z_k}{\pi_k}}$$

[13] Little (2012). Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics (with discussion) *Journal of Official Statistics*, 28: 309-372.

15 The weighting game

- Sample membership indicators:

$$Z_k = \begin{cases} 1, k \in S \\ 0, k \notin S \end{cases}$$

$$E(Z_k) = \pi_k \qquad E(Z_k Z_\ell) = \pi_{k\ell}$$

- The Z_k are random variables; the Y_k are constants
- The **Horvitz-Thompson**, unbiased estimate of T and nearly unbiased of A:

$$\hat{T} = HT[Y_k] = \sum_{k \in S} \frac{Y_k}{\pi_k} = \sum_{k \in U} \frac{Z_k Y_k}{\pi_k}$$

$$\hat{A} = \frac{\sum_{k \in S} \frac{Y_k}{\pi_k}}{\sum_{k \in S} \frac{1}{\pi_k}} = \frac{\sum_{k \in U} \frac{Z_k Y_k}{\pi_k}}{\sum_{k \in U} \frac{Z_k}{\pi_k}}$$

- **Alternatively**, include a flexible function of the πs as a covariate in a regression with the observed Y_k as dependent variable[13]
  - The goal is to make the selection probabilities 'ignorable'

[13] Little (2012). Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics (with discussion) *Journal of Official Statistics*, 28: 309-372.

The good news and the cautions

- If the π_k are correct, the estimator is unbiased w/o needing a model for the Y_ks
- However, in many surveys producing the π_k is complicated, and computing the $\pi_{k\ell}$ is (complicated)^2
  - Non-response, imputation, etc. must be accommodated
- Variance computations are also complicated
- Inferences for non-linear functions on the Ys can be challenging
- Validity depends on good values for the πs, but big data has little or no information on the π_k, let alone the $\pi_{k\ell}$

Generalizing a clinical trial: Internal vs External Suicide Rates

- Pooled clinical trial suicide rates compared to the age-adjusted rates in the nationally representative, Youth Risk Behavior Survey (YRBS)[14].

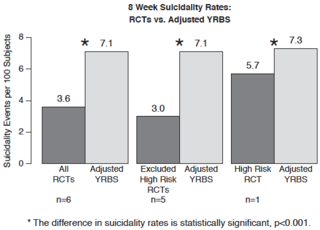

Figure 1. Comparison of the 8-week suicidality rate in the RCT studies (both arms combined) *versus* the age-adjusted YRBS rate. Reading from left to right, the first comparison is of the 6 adolescent MDD RCTs; next is the subset of RCTs that excluded patients at high baseline risk of suicidality; and finally is the one RCT that did not exclude high-risk patients.

- These discrepancies, even after adjustments, highlight the challenges

[14] Greenhouse, Kaizar, Kelleher, Seltman, Gardner (2008). Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. *Statistics in Medicine*, 27: 1801-1813

Generalizing clinical trials and other studies[15,16,17]

- There are three, principal approaches to generalization/transportation:
  - Weighting by sample inclusion propensities
  - Flexible regression modeling or machine learning with propensities as a subset of regressors
    - Applying the model using a target population covariate distribution
  - A combination of the two (double-robustness, targeted MLE)
- Prerequisites for each approach are,
  - Identification of a reference population
  - Measurement of covariates that associate with trial (sample) membership and with treatment (more generally, a relation of interest)
  - The usual ignorability assumptions (hopes)
- The regression approach can proceed with only data from the observed sample, opening the door to progress in the big data context

---

[15] Ackerman, et al. (2019). Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive Behaviors*, 94: 124–132.

[16] Nguyen, et al. (2018). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *Plos one*, e0208795.

[17] Stuart et al. (2018). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on social work practice*, 28: 532–537.

19 Regression using only the observed sample

- Propensities aren't available, but if covariates are available, employ them via flexible regression modeling or machine learning
- Bias can be reduced by building a rich regression model with covariates that
  - Associate with the dependent variable (empirically assessable)
  - Associate with sample inclusion (not empirically accessible)
- You may not know,
  - if the observed covariates associate with selection, but if they do, then RegML will provide at least a partial adjustment for selection effects, and move towards ignorability
  - the target population, but using data bases you can apply the regression structure to a posited population covariate distribution, with a key assumption being that the selection process is applies
- Obtain relevant data on the target population using big data, data melding, ...
- Sensitivity analysis is essential
- Design, collect what you can, especially what you think associates with selection

Gaussian data with informative sample size

$$
\begin{aligned}
Y_k &\sim N\left(\theta_k, \frac{\sigma^2}{n_k}\right) \\
\bar{\theta} &= K^{-1} \sum_k \theta_k \qquad \text{(population mean)} \\
\hat{\bar{\theta}}_{mle} &= \frac{\sum_k n_k Y_k}{\sum_k n_k} \qquad \text{(biased, if } \mathrm{cor}(\theta_k, n_k) \neq 0\text{)} \\
\hat{\bar{\theta}}_{ube} &= \bar{Y} = \frac{1}{K} \sum_k Y_k \quad \text{(unbiased, but higher variance)}
\end{aligned}
$$

Covariate adjusted approach: flexible spline or polynomial in the n_k:

$$
\begin{aligned}
Y_k &= \beta_0 + \mathrm{flexible}(n_k) \\
\hat{\bar{\theta}}_{regr} &= \hat{\beta}_0 + \frac{1}{K} \sum_k \mathrm{flexible}(n_k)
\end{aligned}
$$

Comparisons

- Create n_k with a specified $E(n) = \bar{n}$ and $ratio = \ddot{n}/\bar{n}$
- Produce θ_k with $E(\theta) = 0, V(\theta) = \tau^2$ and various $\rho = \text{cor}(\theta_k, n_k)$,
- Polynomial regression:
  - Select $d \geq 0$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)$
  - $\zeta_k = \sum_{\nu=0}^{d} \beta_\nu n^\nu$
  - $\theta_k =$ the ζ_k adjusted to have mean 0 and variance τ^2
  - Fit using: lm(Y $\sim$ poly(n, d),weights=n)

Results (computations)

- $\ddot{n}$ the harmonic mean, $V(\hat{\hat{\theta}}_{mle})/V(\hat{\hat{\theta}}_{ube}) = \ddot{n}/\bar{n} \leq 1.0$
- $K = 20, \sigma^2 = 20, \bar{n} = 5$
- $\Delta =$ bias of the MLE; $\rho = \mathrm{cor}(\theta_k, n_k)$, true $d = 6$

Column headings are (ρ, Δ)

	$V_{mle}/V_{ube} = 0.325$				$V_{mle}/V_{ube} = 0.714$
Method	(0,0)	(0.41, 0.72)	(0.41, 2.16)	(0.33, 2.16)	(0.41,0.95)
MLE	32	117	791	791	392
Regr $d = 1$	60	103	445	445	155
$d = 2$	77	102	298	298	112
$d = 3$	87	101	208	208	103
$d = 4$	93	100	156	156	101
$d = 5$	96	99	128	127	100
$d = 6$	98	98	98	113	100

$$\boxed{100 \times \frac{MSE}{MSE_{ube}}}$$

Summary

- A well-constructed regression approach is generally effective

Meng's law of large populations[18]

G = population data	N = population size
n = sample size	$f = n/N$, sampling fraction
R = sample inclusion indicator	

$$\text{Discrepancy} = \bar{G}_n - \bar{G}_N = \rho_{R,G} \times \sqrt{\frac{1-f}{f}} \times \sigma_G$$

Data Quality — Data Quantity — Problem Difficulty

$$\text{MSE}_R = E_R\{\rho^2_{R,G}\} \times \left(\frac{1-f}{f}\right)^2 \times \sigma^2_G$$

- E_R is expectation wrt the distribution of R, conditional on $R_+ = n$
- For Simple Random Sampling, $E_R\{\rho^2_{R,G}\} \propto N^{-1}$ and so $\text{MSE}_R = O(n^{-1})$ as it is for many other probability-based sampling plans
- For non-probabilistic sampling MSE_R might not converge to 0 as n increases

[18] Meng (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 Presidential Election. *Annals of Applied Statistics*, 12: 685–726.

Guidance and Challenges

Guidance

- Data Melding:[19] With inputs from a variety of sources, sampling plans, measurement systems, . . .
  - Harmonize inputs to the degree possible
  - Combine over inputs by calibrating biases, and building a (Bayesian) latent structure model (a rosetta stone) to sort out relations[20,21]
- Collect several covariates, especially those that potentially associate with both the target of inference and the selection process, and include flexible functions of them in a regression or use a machine learning approach
- Use administrative records and other databases to help identify reference populations and sampling fractions
- Measure attributes you may not need to meet current study goals, but that can help transport findings to another context
- Conduct aggressive sensitivity analysis

Challenges

- Meng: Information may not increase with sample size; bias will likely persist
- Quantifying variability[22]

[19] Louis TA (1989). Meta Modeling. Section 1.1 'Biometrics,' In, *Challenges for the '90s*. ASA.

[20] Lohr SL, Raghunathan, TE (2017). Combining Survey Data with Other Data Sources. *Statistical Science*, 32: 293–312.

[21] Mugglin and Carlin (1998). Hierarchical modeling in Geographic Information Systems: population interpolation over incompatible zones. *J. of Agricultural, Biological, and Environmental Statistics*, 3: 111-130.

[22] See, Lohr's, Measuring Uncertainty with Multiple Sources of Data

Guidance and Challenges

Guidance

- Data Melding:[19] With inputs from a variety of sources, sampling plans, measurement systems, . . .
  - Harmonize inputs to the degree possible
  - Combine over inputs by calibrating biases, and building a (Bayesian) latent structure model (a rosetta stone) to sort out relations[20,21]
- Collect several covariates, especially those that potentially associate with both the target of inference and the selection process, and include flexible functions of them in a regression or use a machine learning approach
- Use administrative records and other databases to help identify reference populations and sampling fractions
- Measure attributes you may not need to meet current study goals, but that can help transport findings to another context
- Conduct aggressive sensitivity analysis

Challenges

- Meng: Information may not increase with sample size; bias will likely persist
- Quantifying variability[22]

Statistical concepts and techniques are essential for success

---

[19] Louis TA (1989). Meta Modeling. Section 1.1 'Biometrics,' In, *Challenges for the '90s*. ASA.

[20] Lohr SL, Raghunathan, TE (2017). Combining Survey Data with Other Data Sources. *Statistical Science*, 32: 293–312.

[21] Mugglin and Carlin (1998). Hierarchical modeling in Geographic Information Systems: population interpolation over incompatible zones. *J. of Agricultural, Biological, and Environmental Statistics*, 3: 111-130.

[22] See, Lohr's, Measuring Uncertainty with Multiple Sources of Data

#thankyou