

IntermedioSIS – Statistica ed Ambiente – Messina, 22/9/2005

Nonparametric Methods for Sample Surveys of Environmental Populations

Metodi Nonparametrici nell'Inferenza per Popolazioni Finite di carattere Ambientale

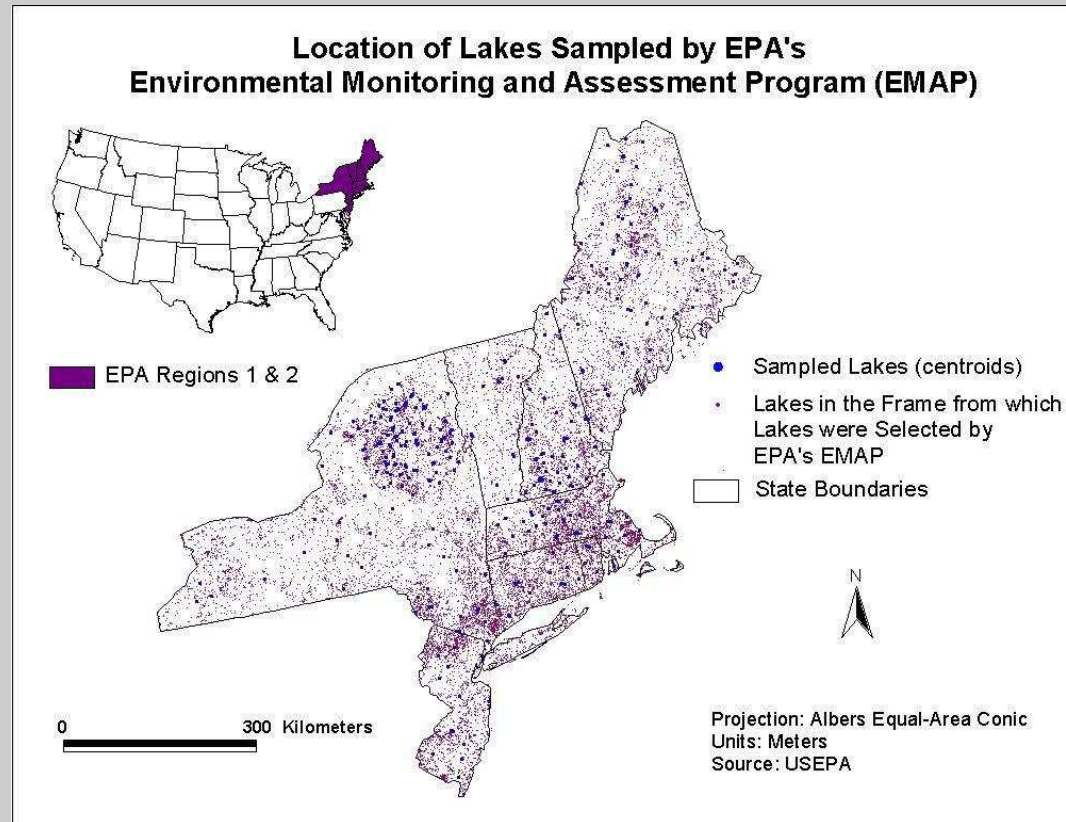
Giorgio E. Montanari & M. Giovanna Ranalli

Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Italy



Problem

Northeastern Lakes – EMAP surveyed ANC in 334 out of 21,026 lakes in the years 1991-1996 (some revisited). How many lakes are acidic/at risk of acidification?



How to deal with it

- It could be determined through Acid Neutralizing Capacity (ANC) thresholds
 - $ANC < 200 \rightarrow$ risk of acidification
 - $ANC < 50 \rightarrow$ high risk of acidification \Rightarrow ESTIMATION OF THE CDF
 - $ANC < 0 \rightarrow$ acidified lake

How to deal with it

- It could be determined through Acid Neutralizing Capacity (ANC) thresholds
 - $ANC < 200 \rightarrow$ risk of acidification
 - $ANC < 50 \rightarrow$ high risk of acidification \Rightarrow ESTIMATION OF THE CDF
 - $ANC < 0 \rightarrow$ acidified lake
- Lakes are selected through a complex design from a frame of lakes \rightarrow finite population approach to get the estimate and the confidence bounds
- Auxiliary information available for each lake in the frame

Formalizing the problem

- $\mathcal{U} = \{u_1, \dots, u_N\}$ is the finite population of lakes labeled by the integers $i = 1, \dots, N$ [$N = 21,026$ in our application];
- y_i is the value taken by the survey variable y [ANC] in unit i [average over revisits];
- $z_i = I(y_i \leq t)$ is the indicator variable whose population mean $F_N(t) = N^{-1} \sum_{i \in \mathcal{U}} z_i$ is the parameter of interest i.e. the CDF at t [$t = 0, 50, 200$];

Formalizing the problem

- $\mathcal{U} = \{u_1, \dots, u_N\}$ is the finite population of lakes labeled by the integers $i = 1, \dots, N$ [$N = 21,026$ in our application];
- y_i is the value taken by the survey variable y [ANC] in unit i [average over revisits];
- $z_i = I(y_i \leq t)$ is the indicator variable whose population mean $F_N(t) = N^{-1} \sum_{i \in \mathcal{U}} z_i$ is the parameter of interest i.e. the CDF at t [$t = 0, 50, 200$];
- s is the sample of size n drawn from \mathcal{U} according to a probabilistic sampling plan with inclusion probabilities π_i and π_{ij} for all $i, j \in \mathcal{U}$ [$n = 334$];
- we have y_i known for $i \in s$; the Hajek estimator for $F_N(t)$ is

$$\hat{F}_H(t) = \frac{\sum_{i \in s} d_i z_i}{\sum_{i \in s} d_i} = \sum_{i \in s} d_i^* z_i$$

with design weights $d_i = 1/\pi_i$ and $d_i^* = d_i / \sum_{i \in s} d_i$.

Auxiliary Information

- The Hajek estimator does not employ AI.
- We have y_i known for all $i \in s$ AND $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{Qi})$ known for all $i \in \mathcal{U}$; in particular for our application

x_{1i} = x -geographical coordinate of the centroid of each lake,

x_{2i} = y -geographical coordinate,

x_{3i} = categorical variable for eco-region (7 levels),

x_{4i} = elevation

- It should be possible to improve the efficiency using AI.



AI for the estimation of a population mean: an overview

- Model based/assisted approach



AI for the estimation of a population mean: an overview

- Model based/assisted approach
- To estimate the population mean of y , consider an assisting model such that $E_{\xi}(y_i) = \mu(\mathbf{x}_i)$ and some design-based estimates $\hat{\mu}_i$:
 - GREG-type estimators $\bar{y}_G = \frac{1}{N} \sum_{i \in \mathcal{U}} \hat{\mu}_i + \frac{1}{N} \sum_{i \in \mathcal{S}} d_i (y_i - \hat{\mu}_i)$.
 - Calibration-type estimators $\bar{y}_C = \frac{1}{N} \sum_{i \in \mathcal{S}} w_i y_i$ with weights that minimize a distance from d_i under the constraints

$$\frac{1}{N} \sum_{i \in \mathcal{S}} w_i \mathbf{x}_i = \frac{1}{N} \sum_{i \in \mathcal{U}} \mathbf{x}_i \quad \text{and/or} \quad \frac{1}{N} \sum_{i \in \mathcal{S}} w_i \hat{\mu}_i = \frac{1}{N} \sum_{i \in \mathcal{U}} \hat{\mu}_i.$$

AI for the estimation of a population mean: an overview

- Model based/assisted approach
- To estimate the population mean of y , consider an assisting model such that $E_{\xi}(y_i) = \mu(\mathbf{x}_i)$ and some design-based estimates $\hat{\mu}_i$:
 - GREG-type estimators $\bar{y}_G = \frac{1}{N} \sum_{i \in \mathcal{U}} \hat{\mu}_i + \frac{1}{N} \sum_{i \in \mathcal{S}} d_i (y_i - \hat{\mu}_i)$.
 - Calibration-type estimators $\bar{y}_C = \frac{1}{N} \sum_{i \in \mathcal{S}} w_i y_i$ with weights that minimize a distance from d_i under the constraints
$$\frac{1}{N} \sum_{i \in \mathcal{S}} w_i \mathbf{x}_i = \frac{1}{N} \sum_{i \in \mathcal{U}} \mathbf{x}_i \quad \text{and/or} \quad \frac{1}{N} \sum_{i \in \mathcal{S}} w_i \hat{\mu}_i = \frac{1}{N} \sum_{i \in \mathcal{U}} \hat{\mu}_i.$$
- Type of model depends on the level of AI available:
 - population totals/means \rightarrow **linear models** \rightarrow GREG / calibration
 - complete AI \rightarrow **nonlinear/generalized linear/nonparametric models** \rightarrow more general GREG / model calibration

Remote sensed AI in nonparametric model-assisted inference for environmental populations

- Local polynomials GREG on a two-stage sample from the National Resource Inventory Erosion Update Survey (Kim et al., 2004)
- Generalized Additive Models GREG on a two-phase sample from a forest inventory in Utah (Opsomer et al., 2005)
- P-splines GREG to estimate ANC mean for the Northeastern lakes survey (Breidt et al., 2005)
- Neural Networks Model Calibration for streams surveyed in the Mid-Atlantic Highlands (Montanari & Ranalli, 2005)

CDF estimation issues

- Straightforward application of these techniques gives

$$\hat{F}_G(t) = \frac{1}{N} \sum_{i \in \mathcal{U}} \hat{z}_i + \frac{1}{N} \sum_{i \in \mathcal{S}} d_i (z_i - \hat{z}_i)$$

with $\hat{z}_i = I(\hat{\mu}_i \leq t)$.

- Alternatively one can use as the auxiliary variable the estimated probability g_i of $y_i \leq t$, i.e.

$$\hat{F}_G(t) = \frac{1}{N} \sum_{i \in \mathcal{U}} g_i + \frac{1}{N} \sum_{i \in \mathcal{S}} d_i (z_i - g_i)$$

- This is not a distribution function and can take values outside $[0, 1]$.

CDF estimation: our approach

- We will employ a Nonparametric Model Calibrated Pseudo-Empirical Maximum Likelihood approach



CDF estimation: our approach

- We will employ a Nonparametric Model Calibrated Pseudo-Empirical Maximum Likelihood approach
- Three main pieces of the jigsaw:
 - Nonparametric regression modeling — N — in particular MARS
 - Model Calibration — MC
 - Pseudo-Empirical Maximum Likelihood — PEML



The PEML piece of the jigsaw

The PEML estimator of the CDF is given by

$$\hat{F}_{\text{PEML}}(t) = \sum_{i \in s} \hat{p}_i z_i$$

The PEML piece of the jigsaw

The PEML estimator of the CDF is given by

$$\hat{F}_{\text{PEML}}(t) = \sum_{i \in s} \hat{p}_i z_i$$

with weights such that

$$\max_{p_i} l(\mathbf{p}) = \sum_{i \in s} d_i \log p_i$$



The PEML piece of the jigsaw

The PEML estimator of the CDF is given by

$$\hat{F}_{\text{PEML}}(t) = \sum_{i \in s} \hat{p}_i z_i$$

with weights such that

$$\max_{p_i} l(\mathbf{p}) = \sum_{i \in s} d_i \log p_i$$

subject to

$$0 < p_i < 1, \quad \sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i g_i = N^{-1} \sum_{i \in \mathcal{U}} g_i$$

where $g_i = g(\mathbf{x}_i)$ is a known function of \mathbf{x}_i (Chen & Sitter, St.Sinica, 1999).

The PEML piece of the jigsaw

The PEML estimator of the CDF is given by

$$\hat{F}_{\text{PEML}}(t) = \sum_{i \in s} \hat{p}_i z_i$$

with weights such that

$$\max_{p_i} l(\mathbf{p}) = \sum_{i \in s} d_i \log p_i$$

subject to

$$0 < p_i < 1, \quad \sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i g_i = N^{-1} \sum_{i \in \mathcal{U}} g_i$$

where $g_i = g(\mathbf{x}_i)$ is a known function of \mathbf{x}_i (Chen & Sitter, St.Sinica, 1999).

- Asymptotically equivalent to a GREG if $g(\mathbf{x}_i) = \mathbf{x}_i$. It looks a lot like calibration!
- (Newton-Raphson type) algorithms to find the solution (Chen et al., Biom.ka, 2002).
- It is equivalent to maximize $l_n(\mathbf{p}) = n^* \sum_{i \in s} d_i^* \log p_i$; useful for CI derivation.

The MC piece of the jigsaw – choice of g_i 's

For a given value t_0 the *optimal* choice (in the sense of Wu, Biometrika, 2003) is

$$g_i = E_{\xi}(z_i | \mathbf{x}_i) = P(y_i \leq t_0 | \mathbf{x}_i),$$

where ξ is a superpopulation (assisting) model.



The MC piece of the jigsaw – choice of g_i 's

For a given value t_0 the *optimal* choice (in the sense of Wu, Biometrika, 2003) is

$$g_i = E_{\xi}(z_i|\mathbf{x}_i) = P(y_i \leq t_0|\mathbf{x}_i),$$

where ξ is a superpopulation (assisting) model.

→ Given the (binary) nature of the response variable z_i , a natural candidate for ξ is the logistic regression model (Chen & Wu, St.Sinica, 2002)

$$\log \left(\frac{g_i}{1 - g_i} \right) = \mathbf{x}_i \boldsymbol{\beta}.$$



The N piece of the jigsaw – a more flexible model



The N piece of the jigsaw – a more flexible model

- Evidence from other studies on ANC in the NorthEastern lakes (Opsomer et al., WP, 2004; Breidt et al., WP, 2005) shows the need for a more complex model in the auxiliary variables.



The N piece of the jigsaw – a more flexible model

- Evidence from other studies on ANC in the NorthEastern lakes (Opsomer et al., WP, 2004; Breidt et al., WP, 2005) shows the need for a more complex model in the auxiliary variables.
- We therefore extended the logistic model formulation to accommodate non-linear relationships:

$$\log \left(\frac{g_i}{1 - g_i} \right) = \mu(\mathbf{x}_i),$$

where $\mu(\cdot)$ is an unknown function. Nonparametric techniques can be used to obtain (design-based) estimates of $\mu(\cdot)$.

- We will use Multi Adaptive Regression Splines – MARS – in the application.

Confidence intervals

- Instead of using a normal limiting distribution of the estimators, we exploit the pseudo-empirical likelihood nature of the derivation to obtain a CI for $F_N(t)$ at $t = \tilde{t}$. Let's recall it **9**

Confidence intervals

- Instead of using a normal limiting distribution of the estimators, we exploit the pseudo-empirical likelihood nature of the derivation to obtain a CI for $F_N(t)$ at $t = \tilde{t}$. Let's recall it **9**
- Let \tilde{p}_i be the value of p_i that maximizes $l_n(\mathbf{p})$ subject to

$$0 < p_i < 1, \quad \sum_{i \in S} p_i = 1, \quad \sum_{i \in S} p_i g_i = N^{-1} \sum_{i \in \mathcal{U}} g_i, \quad \sum_{i \in S} p_i z_i = \theta,$$

for $z_i = I(y_i \leq \tilde{t})$ and a fixed θ .

Confidence intervals

- Instead of using a normal limiting distribution of the estimators, we exploit the pseudo-empirical likelihood nature of the derivation to obtain a CI for $F_N(t)$ at $t = \tilde{t}$. Let's recall it 9
- Let \tilde{p}_i be the value of p_i that maximizes $l_n(\mathbf{p})$ subject to

$$0 < p_i < 1, \quad \sum_{i \in S} p_i = 1, \quad \sum_{i \in S} p_i g_i = N^{-1} \sum_{i \in \mathcal{U}} g_i, \quad \sum_{i \in S} p_i z_i = \theta,$$

for $z_i = I(y_i \leq \tilde{t})$ and a fixed θ .

- It can be proved (Wu & Rao, 2004) that the ratio statistics

$$r_n(\theta) = -2\{l_n(\tilde{\mathbf{p}}) - l_n(\hat{\mathbf{p}})\} \xrightarrow{d} \chi_1^2,$$

so that a $(1 - \alpha)$ PEML CI for $F_N(\tilde{t})$ is given by the set

$$\{\theta | r_n(\theta) < \chi_1^2(\alpha)\}.$$

Application to ANC CDF estimation in the NE lakes

- Recall that we have $n = 334$ possibly averaged measurements of ANC in lakes surveyed from a frame of $N = 21,026$.
- The Hajek estimator and the PEML estimator that uses MARS of $F_N(t)$ have been computed at $t = (0, 50, 200)$ and at a 1000 value grid.
- The CI computation has been conducted adapting some R functions provided in Wu (WP, 2005).



Modeling issues

- MARS has been used to approximate the unknown function $\mu(\mathbf{x}_i)$ in the nonparametric logistic model.
- Recall the optimal choice of g_i **10**: it depends on t_0 → no g_i with a fixed t_0 can be uniformly optimal for $F_N(t)$ for all values of t .



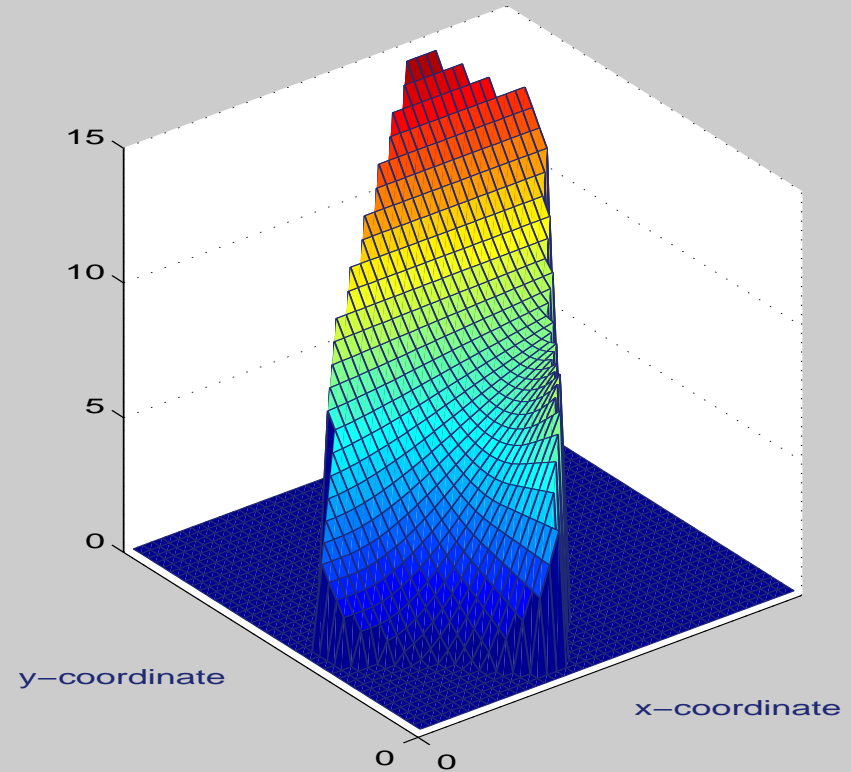
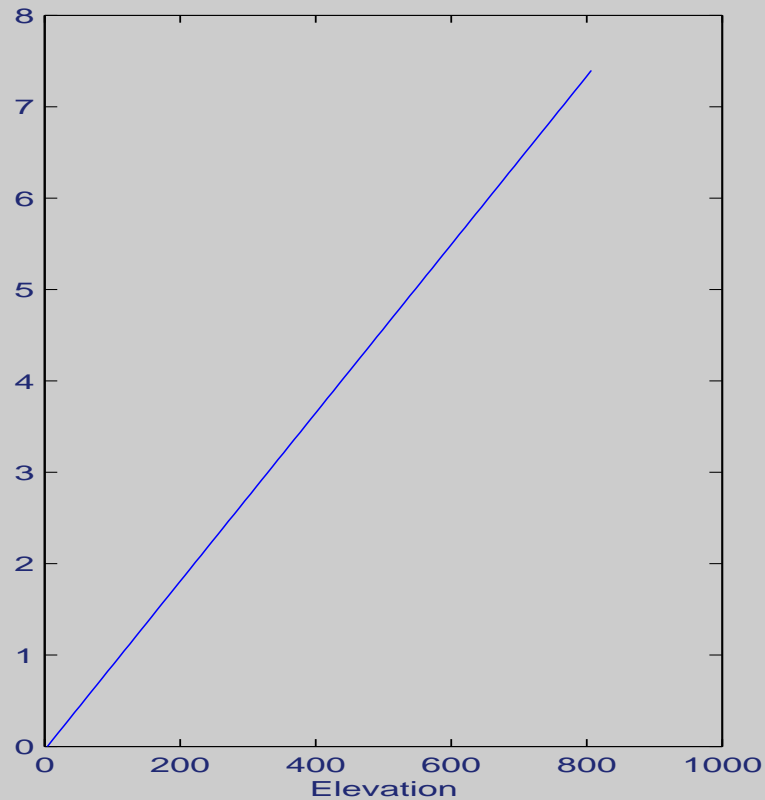
Modeling issues

- MARS has been used to approximate the unknown function $\mu(\mathbf{x}_i)$ in the nonparametric logistic model.
- Recall the optimal choice of g_i **10**: it depends on t_0 → no g_i with a fixed t_0 can be uniformly optimal for $F_N(t)$ for all values of t .
- The model has been fitted at $t_0 = 200$, i.e. the estimated g_i to use in the MCP EML procedure have been obtained only for the model that relates $I(y_i \leq 200)$ to the \mathbf{x}_i 's and then used for all t 's. This is not optimal, but guarantees the achievement of a genuine CDF.

Modeling issues

- MARS has been used to approximate the unknown function $\mu(\mathbf{x}_i)$ in the nonparametric logistic model.
- Recall the optimal choice of g_i **10**: it depends on $t_0 \rightarrow$ no g_i with a fixed t_0 can be uniformly optimal for $F_N(t)$ for all values of t .
- The model has been fitted at $t_0 = 200$, i.e. the estimated g_i to use in the MCP EML procedure have been obtained only for the model that relates $I(y_i \leq 200)$ to the \mathbf{x}_i 's and then used for all t 's. This is not optimal, but guarantees the achievement of a genuine CDF.
- The generalized cross validation criterion considered in Friedman (Annals, 1991) suggested the use of 15 basis functions; more interestingly, all variables turned out to be significant according to this criterion: elevation enters as an additive variable (no interactions with other variables), while the x and y -geographical coordinates show a significant interaction.

Curves and surfaces estimated by MARS



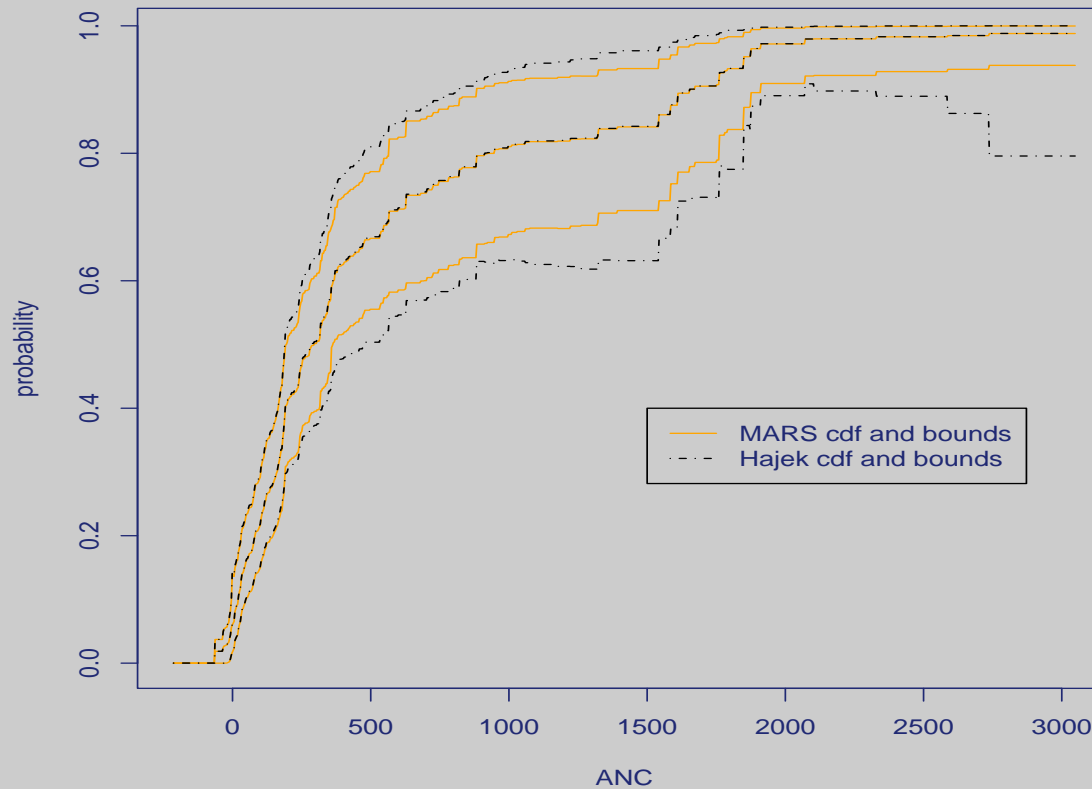
The vertical axis in each plot shows the contribution of each variable to the whole smooth predictor $\hat{\mu}_i$; since the locations of the plotted functions are arbitrary, they are all translated to have zero minimum value.

CDF estimates at the three thresholds

t	Hajek		MARS	
	$\hat{F}_H(t)$	95% CI	$\hat{F}_{\text{MARS}}(t)$	95% CI
0	0.060	(0.017; 0.143)	0.060	(0.017; 0.140)
50	0.164	(0.104; 0.238)	0.162	(0.103; 0.234)
200	0.411	(0.301; 0.527)	0.408	(0.311; 0.505)

Cdf estimates at $t = 0, 50, 200$ and relative 95% confidence intervals by the Hajek and Mars estimators. The average length of the confidence intervals is 0.162 with $\hat{F}_H(t)$ and 0.149 with $\hat{F}_{\text{MARS}}(t)$.

CDF estimate and CI bounds at 1000 value grid



The average length of the confidence intervals is 0.209 with $\hat{F}_H(t)$ and 0.149 with $\hat{F}_{\text{MARS}}(t)$.

Wrap up

- Nonparametric Model Calibration has been applied to PEML to allow for more flexible models that use complete AI.
- The PEML framework is particularly suitable for CDF estimation: it provides weights with desirable properties and allows for more efficient estimation of confidence bounds.

Wrap up

- Nonparametric Model Calibration has been applied to PEML to allow for more flexible models that use complete AI.
- The PEML framework is particularly suitable for CDF estimation: it provides weights with desirable properties and allows for more efficient estimation of confidence bounds.
- An application to ANC CDF estimation in NE lakes has been conducted: the use of AI has shown improvements in estimated efficiency in the form of average confidence bounds 40% wider for the Hajek estimator.
- The application of MARS provides sensible modeling results without unduly *a priori* assumptions on the way the auxiliary variables enter the model.

To do list

- Study the dependence of the final estimates on the choice of the t_0 value for which we fit the model and obtain the weights; how much efficiency is lost?



To do list

- Study the dependance of the final estimates on the choice of the t_0 value for which we fit the model and obtain the weights; how much efficiency is lost?
- ...then find some sensible guidelines to choose it (!)

To do list

- Study the dependence of the final estimates on the choice of the t_0 value for which we fit the model and obtain the weights; how much efficiency is lost?
- ...then find some sensible guidelines to choose it (!)
- Set up a more thorough simulation study to understand when the PEML CI estimate is better than the simple normal theory one (the PEML one can be *lengthy* for t 's faraway from t_0).

Essential bibliography and acknowledgments

Breidt F.J., Opsomer J.D., Johnson A.A., Ranalli M.G. (2005) Semiparametric model-assisted estimation for natural resources surveys, *WP*.

Chen J. & Sitter R.R. (1999) A pseudo-empirical likelihood approach to the effective use of auxiliary information in complex surveys, *St. Sinica*, 9, 385–406.

Chen J., Sitter R.R. & Wu C. (2002) Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika*, 89, 1, 230–237.

Chen J. & Wu C. (2002) Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method, *St. Sinica*, 12, 1223–1239.

Opsomer J.D., Breidt F.J., Claeskens G., Kauermann G., Ranalli M.G. (2004) Nonparametric small area estimation using penalized spline regression, *ASA Proceedings on Survey Research Methods*

Wu C. (2003) Optimal calibration estimators in survey sampling, *Biometrika*, 90, 937–951.

Wu C. & Rao J. (2004) Pseudo empirical likelihood ratio confidence intervals for complex surveys, *WP*.

The work reported here was developed by the second author under the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of the presenter and STARMAP. EPA does not endorse any products or commercial services mentioned in this presentation.