Predictive Inference of Machine Learning for Official Statistics

Li-Chun Zhang

University of Southampton (L.Zhang@soton.ac.uk)
Statistisk sentralbyrå

Outline

We are concerned with official statistics that are defined as **descriptive summaries** of finite populations

We shall consider ML models as **statistical algorithms**, with specified input data and algorithmically generated output (as estimates or predictions), which bypasses the ideal existence (or knowledge) of 'true data model'

> *How to make valid **predictive inference** (i.e. about descriptive target) **using arbitrary ML models** that are often patently not 'true'?*

We shall highlight and discuss

- **assumption-lean** predictive inference using ML
- supervised learning, **transfer learning**
- future developments (if time allows)

Assumption-learn predictive inference

I-1. Use what is known instead of what is assumed, i.e. use the sampling design if it exists.

Basic idea of design-based model-assisted estimation...
However, plug-in residual-correction (the GREG "recipe") requires justification for *arbitrary* models

In case model-assisted estimator is *not* design-unbiased, common to seek *design-consistent asymptotic inference*

- hypothetical sequence of increasing populations
- tailored arguments for specific algorithmic ML
  e.g. Dagdoug et al. (2023) for random forest
- still may not hold in given finite-sample setting
  Smith (1994): the "asymptotic notion of consistency" vs. the given population as "a real entity"

Assumption-learn predictive inference

I-2. Make Neyman-Fisher consistent inference, resort to asymptotic inference if necessary.

Neyman-Fisher consistency (Zhang et al., 2025)

- estimator $g(t_1, ..., t_k)$ is called "consistent" for $g(T_1, ..., T_k)$, if $t_1, ..., t_k$ are unbiased of the population totals $T_1, ..., T_k$ (Fisher, 1956)
- an interval estimator of "a collective character ... of a population" is said to be "consistent", if it achieves the designated level of coverage (Neyman, 1934)

Recently, design-based predictive inference

- pq-design, subsampling Rao-Blackwellisation (SRB) (Sanguiao-Sande & Zhang, 2021), a new "recipe"
- prediction estimation (Zhang et al., 2025)
  NB. at both population and unit levels
- prediction intervals (Zhang & Sanguiao-Sande, 2025)
  NB. to be contrasted to conformal prediction

Assumption-learn predictive inference

II. Make assumption-lean model-based inference, e.g. assume exchangeable joint data† distributions but not ML model that is patently not 'true'.

Model-based inference may be necessary
e.g. combining multiple non-survey big data sources

Generally, what may be considered as 'assumption-lean' model-based inference varies in different situations!

Conformal intervals given IID or exchangeability,
'lean' by assumed analogy to design-based inference

- *Quasi-randomisation*: absence of probability sampling, but IID (or exchangeability) is *assumed* for necessary and assumption-lean model-based inference
- IID if finite-population with-replacement sampling
- exchangeable sample observations† if SRS $^{without}_{replacement}$

Assumption-learn predictive inference

III. Use 'prior' information by transfer learning (TL), when unbiased supervised learning (either by design or model) is too noisy or not possible at all.

E.g. *inductive* transfer learning (Pan and Yang, 2009) given "labels" in both *target* and *source* populations

Zhang (2025): utilise observations in source population as 'empirical' *prior information* in *non-Bayesian* ways

- transfer learning estimation: biased but smaller MSE
- useful if supervised learning is too noisy or impossible
- leaner than Bayesian prior distribution of unobservable e.g. design-based TL (if not Neyman-Fisher consistent)

TL-paradigm has many applications in official statistics

- TL-HT estimator, ensemble SAE, flash estimation...

Design-based predictive inference supervised learning

Design-based predictive inference

Population U and associated $y_U = \{y_i : i \in U\}$

Element	i_1	i_2	i_3	i_4
Value y_i	1	2	3	6

Simple random sampling without replacement, $n = 2$

Sample s	(i_1, i_2)	(i_1, i_3)	(i_1, i_4)	(i_2, i_3)	(i_2, i_4)	(i_3, i_4)
Sample mean $\bar{y}_s$	1.5	2	3.5	2.5	4	4.5
Out-of-sample mean $\bar{y}_R(s)$	4.5	4	2.5	3.5	2	1.5
Unknown y_k for $k \notin s$	$y_3 = 3$	$y_2 = 2$	$y_2 = 2$	$y_1 = 1$	$y_1 = 1$	$y_1 = 1$
	$y_4 = 6$	$y_4 = 6$	$y_3 = 3$	$y_4 = 6$	$y_3 = 3$	$y_2 = 2$
$(y_k - \hat{y}_k)^2 = (y_k - \bar{y}_s)^2$	2.25	0	2.25	2.25	9	12.25
	20.25	16	0.25	12.25	1	6.25
$\bar{D}_R(s) = \frac{1}{2}\sum_{k \notin s}(y_k - \bar{y}_s)^2$	11.25	8	1.25	7.25	5	9.25

Fundamental idea: treat $\{(x_i, y_i) : i \in U\}$ as *constants*, define out-of-sample *prediction* and its *performance* only with respect to $s \sim p(s)$

Inference setting: pq-design

pq-**design** of (s, s_1):

- **sample** $s \sim p(s)$ from given finite population U
- **training set** $s_1 \sim q(s_1 \mid s)$ from realised sample s
  test (or holdout) set $s_2 = s \backslash s_1$
  just like sample-split for cross-validation in ML

Well-defined pq-design:

$$q(s_1 \mid s)p(s) = f(s_1)f(s \mid s_1)$$

$$\pi_{2i} = \Pr(i \in s_2 \mid s_1) \;\Rightarrow\; E_s\Big(\sum_{i \in s_2} \tfrac{c_i}{\pi_{2i}} \mid s_1\Big) = \sum_{i \in U \backslash s_1} c_i$$

$$E_p\big(E_q(\cdot \mid s)\big) = E_{s_1}\big(E_s(\cdot \mid s_1)\big)$$

In particular, **SRB** refers to the operation $E_q(\cdot \mid s)$
given minimal sufficient statistic (Godambe and Joshi, 1965)

Inference method: SRB

Example: Twice-SRS pq-design of s from U and s_1 from s

$$\pi_{2i} = n_2/(N - n_1)$$

where n_1 is the size of s_1 and $n_2 = n - n_1$ that of s_2
Let $\mu(x, s_1) = \bar{y}_{s_1}$ by subsample s_1, and $\mu(x, s) = \bar{y}_s$ by s
By SRB, we obtain the **SRB-predictor**

$$\bar{\mu}(x, s) = E_q(\bar{y}_{s_1} \mid s) = \bar{y}_s = \mu(x, s)$$

Distinct notation $\bar{\mu}(x, s)$ to full-sample once-trained $\mu(x, s)$ since they may differ generally, e.g.

$$\mu(x, s) = x^\top (\sum_{i \in s} x_i x_i^\top)^{-1} (\sum_{i \in s} x_i y_i)$$

$$\bar{\mu}(x, s) = x^\top E_q\Big((\sum_{i \in s_1} x_i x_i^\top)^{-1} (\sum_{i \in s_1} x_i y_i) \mid s \Big)$$

Population estimation

Design-unbiased estimation using *arbitrary* ML model

$$\hat{Y}_{1M} = \sum_{i \in s_1} y_i + \left\{ \sum_{i \in U \setminus s_1} \mu(x_i, s_1) + \sum_{i \in s_2} \frac{y_i - \mu(x_i, s_1)}{\pi_{2i}} \right\}$$

i.e. plug-in **error** (not **residual**) of $\mu(x_i, s_1)$, such that

$$E_s(\hat{Y}_{1M} \mid s_1) = Y$$

conditional on s_1 under the pq-design. Applying SRB to $\hat{Y}_{1M}$ yields the **SRB estimator**

$$\hat{Y}_M = E_q(\hat{Y}_{1M}|s)$$

Combining three classic ideas of statistical inference

- model-assisted estimation in survey sampling
- cross-validation for ~~residual~~ error estimation in ML
- Rao-Blackwellisation for efficiency improvement

Prediction estimation

Direct out-of-sample prediction without error correction:

$$\hat{Y}_1^* = \sum_{i \in s} y_i + \sum_{j \in U \setminus s} \mu(x_j, s_1)$$

$$\hat{Y}^{RB} = \sum_{i \in s} y_i + \sum_{j \in U \setminus s} \bar{\mu}(x_j, s)$$

Incl. GREG/model-calibrated estimator, SRB-estimator, and HT-estimator, with $x_i = \pi_i N/n$ and $\pi_i = \Pr(i \in s)$,

$$\mu_{HT}(x_j, s) = x_j \beta_s + \tfrac{1}{N-n} \sum_{i \in s} (x_i \beta_s - y_i) \qquad \left[\beta_s = \tfrac{1}{n} \sum_{i \in s} \tfrac{y_i}{x_i}\right]$$

Out-of-sample SRB-prediction error over $R = U \setminus s$

$$\begin{cases} \left\{ \sum_{i \in R} \bar{\mu}(x_i, s) - \sum_{i \in R} y_i \right\}^2 & \text{STE for population} \\ \sum_{i \in R} \{\bar{\mu}(x_i, s) - y_i\}^2 & \text{TSE for units} \end{cases}$$

Neyman-Fisher consistent inference

Zhang et al. (2025): design-unbiased estimation of

$$\mathrm{MSE}(\hat{Y}^{RB}) = E_p\big(\mathrm{STE}(\bar{\mu})\big)$$

$$\tau(\bar{\mu}) = \tfrac{1}{N-n} E_p\big(\mathrm{TSE}(\bar{\mu})\big)$$

- Prediction problem by definition, ..., property evaluated w.r.t. $s \sim p(s)$; contrasting outlook to model-assisted estimation of fixed population parameters
- *Neyman-Fisher consistent* bias/MSE estimation, given arbitrary ML model $\mu(x)$; *fully design-based* inference in contrast to *design-consistent asymptotic* inference
- Design-theoretical basis for individual-level prediction (estimation) as another distinction to model-assisted theory only applicable for population estimation

Prediction intervals, unit level

Design-based subsample or SRB prediction intervals

$$\begin{aligned}\mathcal{C}(x, s_1) &= \big(\mu_L(x, s_1),\ \mu_U(x, s_1)\big) \\ \bar{\mathcal{C}}(x, s) &= \big(\bar{\mu}_L(x, s),\ \bar{\mu}_U(x, s)\big) \\ &= \big(E_q\{\mu_L(x, s_1) \mid s\},\ E_q\{\mu_U(x, s_1) \mid s\}\big)\end{aligned}$$

Zhang & Sanguiao-Sande (2025) define

realised coverage $\alpha(s_1) = \dfrac{1}{N - n_1} \sum_{i \notin s_1} \mathbb{I}(y_i \in \mathcal{C}(x_i, s_1))$

expected coverage $\bar{\alpha} = E_{pq}\big(\alpha(s_1)\big)$

with design-unbiased **SRB coverage-estimator**

$$\hat{\bar{\alpha}} = E_q(\hat{\alpha}(s_1))$$

NB. no need to require negligible sampling fractions

Prediction intervals, unit level

Expected coverage of $\bar{\mathcal{C}}(x, s)$, average $\hat{\bar{\alpha}}(s)$, API population

$(N = 6194)$	$n = 100$		$n = 2000$		$n = 4000$	
ϕ	$\bar{\mathcal{C}}(x, s)$	$\hat{\bar{\alpha}}(s)$	$\bar{\mathcal{C}}(x, s)$	$\hat{\bar{\alpha}}(s)$	$\bar{\mathcal{C}}(x, s)$	$\hat{\bar{\alpha}}(s)$
1.5	0.875	0.869	0.880	0.880	0.880	0.880
2.0	0.948	0.944	0.952	0.951	0.952	0.952
2.5	0.978	0.976	0.980	0.980	0.980	0.980

- Marginal intervals, coverage over all out-of-sample units
  NB. Barber et al. (2021): no conformal method can guarantee the realised coverage of any particular out-of-sample unit.
- SRB interval $\bar{\mathcal{C}}(x, s)$ more efficient than $\mathcal{C}(x, s_1)$
  e.g. width of $\mathcal{C}(x, s_1)$ varies with s_1 but not $\bar{\mathcal{C}}(x, s)$
  NB. split-conformal interval inefficient in a similar sense
- Conformal theory focus on pre-specified coverage; design-based theory here aims to improve the efficiency of both **interval** and **coverage** estimation

Prediction intervals, population level

Out-of-sample population mean $\bar{Y}_R(s)$, or $\mathbb{I}(\bar{Y}_R(s) \in \mathcal{C}(s))$, cannot have *exchangeable* in-sample observations
NB. unlike y_j or $\mathbb{I}(y_j \in \mathcal{C}_j(s))$ at unit-level, for $j \in R = U \backslash s$
NB. problem also for design-based prediction intervals
Design-consistent prediction intervals of $\bar{Y}_R(s)$

- pre-trained predictor $\mu(x)$ unaffected by $s \sim p(s)$
- SRB-predictor $\bar{\mu}(x, s)$ based on $s \sim p(s)$

(Zhang & Sanguiao-Sande, 2025), such that

$$\lim_{N \to \infty} \Pr\left(\bar{Y}_R(s) \in \mathcal{C}_{\alpha,\mu}(s)\right) = \alpha$$

asymptotically, as $n, N \to \infty$, $\frac{n}{N} \to \lambda \in (0, 1)$, $\frac{n_1}{n} \to \lambda_1 \in (0, 1)$
Key condition HT-CLT: the HT-estimator satisfies CLT
NB. property of populations & sampling designs, not ML

Prediction intervals, population level

Coverage of $\mathcal{C}_{\alpha,\mu}(s)$ with $\mu(x_i) = x_i$, API population

	Sample size n of SRS							
α	16	24	32	48	64	108	256	512
0.90	0.88	0.89	0.89	0.89	0.89	0.89	0.90	0.90
0.95	0.93	0.94	0.94	0.94	0.94	0.95	0.95	0.95
0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99

Coverage of $\mathcal{C}_{\alpha,\mu}(s_1)$ or $\bar{\mathcal{C}}_{\alpha,\mu}(s)$ with $\mu(x) = \beta_0 + x\beta_1$

	$n = 50$		$n = 100$		$n = 200$		$n = 300$	
α	$\mathcal{C}_{\alpha,\mu}(s_1)$	$\bar{\mathcal{C}}_{\alpha,\mu}(s)$	$\mathcal{C}_{\alpha,\mu}(s_1)$	$\bar{\mathcal{C}}_{\alpha,\mu}(s)$	$\mathcal{C}_{\alpha,\mu}(s_1)$	$\bar{\mathcal{C}}_{\alpha,\mu}(s)$	$\mathcal{C}_{\alpha,\mu}(s_1)$	$\bar{\mathcal{C}}_{\alpha,\mu}(s)$
0.90	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.90
0.95	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.95
0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99

Given HT-CLT, ML predictor $\mu(x)$, $\mu(x, s_1)$ or $\bar{\mu}(x, s)$ should not cause unbounded out-of-sample prediction errors

Transfer learning
non-Bayesian paradigm given prior information

Transfer learning

Pan and Yang (2009): Given a *source* domain D_S and learning task T_S, a *target* domain D and learning task T, **transfer learning (TL)** aims to help improve the learning of the target prediction function $f_T(\cdot)$ in D using the knowledge in D_S and T_S, where $D_S \neq D$, or $T_S \neq T$.

Some settings of transfer learning or supervised learning

Labels in (target, source) population			
(Yes, Yes)	(No, Yes)	(No, No)	(Yes, –)
Inductive	Transductive	Unsupervised	Non-target labels as features
Transfer learning			Supervised learning

NB. "population" instead of "domain"

E.g. target parameter β given source estimate $\hat{\theta}$, obtain

$$\hat{\beta}^{TL} = \arg\min_{\beta} \left(\sum_{i \in s} \{y_i - \mu(x_i; \beta)\}^2 + \gamma \|\beta - \hat{\theta}\|_2 \right)$$

NB. regularised loss; not using $\hat{\theta}$ in $\mu(x)$, or via $\pi(\beta)$

Non-Bayesian paradigm of prior information

Zhang (2025):

- distinct population distribution by parameter θ
- source population observations $\mapsto$ source estimate $\hat{\theta}$
- $\hat{\theta}$ as *empirical* prior information for target β
  e.g. the two should not be too far from each other
- used in a non-Bayesian manner for estimating β
  e.g. regularisation towards $\hat{\theta}$
  instead of assuming $\beta \sim N(\hat{\theta}, \sigma^2)$
- transfer learning as a leaner non-Bayesian paradigm of prior information, for many situations where unbiased supervised learning is either too noisy or impossible

NB. Bayesian priors in the form of distributions

TL-HT estimation

Population U of size N, probability sample s of size n
However, n too small, so that $V_p(\hat{\theta}_{HT}) = V_p(\hat{\bar{y}}_U)$ too large
Suppose proxy x_k for $k \in B$ in a privately-held source B
Access to $\tilde{\theta} = \bar{x}_B$, and $\tilde{\theta} \approx \theta$ but $B \neq U$, $x_k \neq y_k$ generally

How to improve HT with empirical prior $\bar{x}_B$?

- consider

$$\hat{\theta}_{HT} = \frac{1}{n} \sum_{k \in s} \frac{n y_k}{N \pi_k} = \frac{1}{n} \sum_{k \in s} \frac{y_k}{x_k}$$

  by

$$\min_{\theta} \sum_{k \in s} (y_k - x_k \theta)^2 / x_k^2$$

- regularised loss

$$\Delta(\theta; \gamma) = \frac{1}{n} \sum_{k \in s} (y_k - x_k \theta)^2 / x_k^2 + \gamma (\theta - \tilde{\theta})^2$$

$$\Rightarrow \quad \theta(\gamma) = \frac{1}{1+\gamma} \hat{\theta}_{HT} + \frac{\gamma}{1+\gamma} \tilde{\theta}$$

TL-HT estimation

We recover $\hat{\theta}_{HT}$ if $\gamma = 0$, or $\tilde{\theta}$ if $\gamma = \infty$
Minimising design-based MSE of $\theta(\gamma)$ yields

$$\gamma = \frac{(\tilde{\theta} - \theta)^2}{V(\hat{\theta}_{HT}) + (\tilde{\theta} - \theta)^2}$$

TL-HT estimator follows from replacing γ by $\hat{\gamma}$

- TL-HT is assumption-lean, since it is fully design-based (though not Neyman-Fisher consistent)
- non-Bayesian use of empirical prior information can help to reduce design-based MSE
- see MNO-MINDS (2025, Ch. 9) for combining Tourist Survey with aggregated Mobile Network Operator data to estimate inbound tourist trips

TL for ensemble estimation

Areas (or domains) $k = 1, ..., m$
In-sample $\mathcal{D}_1 = \{k : n_k > 0\}$, out-of-sample $\mathcal{D}_0 = \{k : n_0 = 0\}$
Direct estimator $\hat{Y}_k$ for $k \in \mathcal{D}_1$, and $\hat{Y}$ for $Y = \sum_{k=1}^{m} Y_k$
$\hat{Y}$ planned, $\hat{Y}_k$ too noisy if $k \in \mathcal{D}_1$, or non-existent if $k \in \mathcal{D}_0$
Common with model-based estimation, e.g.

$$y_k = \log \hat{Y}_k = z_k^\top \beta + v_k + e_k$$

Given assumptions of function $E(y_k)$, random effects v_k:

EBLUP of Y_k for $k \in \mathcal{D}_1$, synthetic $\hat{\mu}_k = z_k^\top \hat{\beta}$ for $k \in \mathcal{D}_0$

For TL: $\exists X_k$ as proxy of Y_k, $\forall k$; target *ensemble parameter*

$$\theta = (\theta_1, ..., \theta_m)^\top \quad \text{where} \quad \theta_k = \frac{Y_k}{Y} \quad \text{and} \quad \sum_{k=1}^{m} \theta_k = 1$$

source ensemble parameter

$$q = (q_1, ..., q_m)^\top \quad \text{where} \quad q_k = \frac{X_k}{X} \quad \text{and} \quad \sum_{k=1}^{m} q_k = 1$$

Inductive learning for $\mathcal{D}_1$

For re-standardised (θ_k, q_k) within $\mathcal{D}_1$, consider

$$\Delta(\mathcal{D}_1;\gamma) = -\sum_{k\in\mathcal{D}_1} \hat{Y}_k \log\theta_k + \gamma \sum_{k\in\mathcal{D}_1} X_k\big(\log q_k - \log\theta_k\big) + \lambda\Big(\sum_{k\in\mathcal{D}_1}\theta_k - 1\Big)$$

Given non-trivial γ, **TL-estimator**

$$\dot{\theta}_k = wp_k + (1-w)q_k \qquad \text{and} \qquad w = \frac{\hat{Y}(\mathcal{D}_1)/X(\mathcal{D}_1)}{\gamma + \hat{Y}(\mathcal{D}_1)/X(\mathcal{D}_1)}$$

Obtain w by minimising total MSE

$$\text{TMSE}(w) = E\Big(\sum_{k\in\mathcal{D}_1}(\dot{\theta}_k - \theta_k)^2\Big) = w^2\sum_{k\in\mathcal{D}_1} V(p_k) + (1-w)^2\sum_{k\in\mathcal{D}_1}(q_k-\theta_k)^2$$

- design-based, despite semblance to James-Stein (1961)
- generally, allow proxy $\tilde{\theta}_k = g(z_1, ..., z_{m_1};\beta)$, regardless X_k approximate TMSE given $\hat{\beta}$ (not β), still design-based NB. EBLUP-inference assumes 'true' function $E(y_k)$...
- TL can be more efficient than EBLUP (Tano & Zhang, 2025)

Transductive learning for $\mathcal{D}_0$

Use $\mathcal{D}_1$ as source population for $\mathcal{D}_0$
Consider a *pseudo* random effect $\dot{v}_k$ for $k \in \mathcal{D}_0$, such that

$$\dot{\theta}_k = \hat{\mu}_k + \dot{v}_k$$

Minimise $$L(\mathcal{D}_0) = \sum_{k \in \mathcal{D}_0} v_k^2 + \alpha \sum_{k \in \mathcal{D}_0} (\hat{\mu}_k + v_k - q_k)^2$$

$$\Rightarrow \quad v_k(\phi) = \phi \hat{a}_k \quad \text{where} \quad \phi = \tfrac{\alpha}{1+\alpha} \quad \text{and} \quad \hat{a}_k = q_k - \hat{\mu}_k$$

To choose ϕ, notice the following in source population:

$$\text{TMSE}(\phi) = \sum_{k \in \mathcal{D}_1} (\mu_k + \phi a_k - \theta_k)^2$$

$$\Rightarrow \quad \phi = \Big(\sum_{k \in \mathcal{D}_1} a_k^2 \Big)^{-1} \sum_{k \in \mathcal{D}_1} (\theta_k - \mu_k) a_k$$

Transductive $\dot{\theta}_k$ for $k \in \mathcal{D}_0$, given $\hat{\phi}$ from $\mathcal{D}_1$ in addition

Transductive learning for $\mathcal{D}_0$

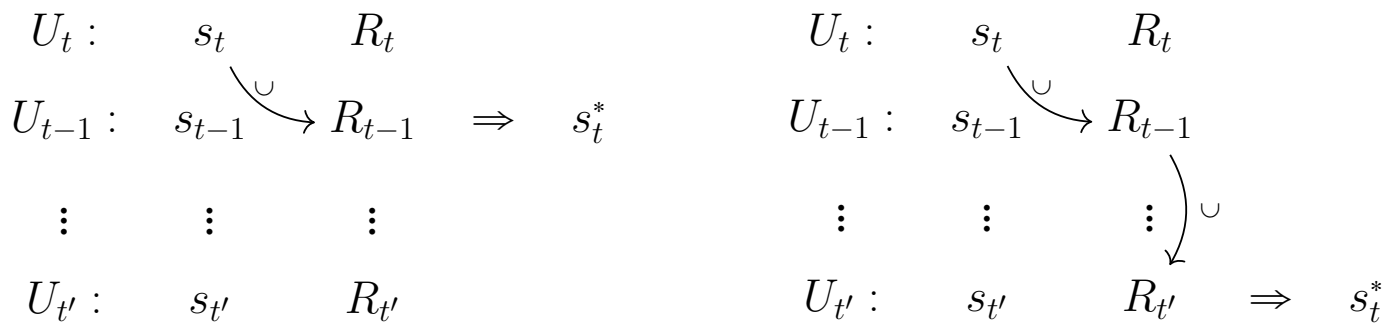
Tuoto and Zhang (2025):

- To avoid 'circular' use of $q_k = X_k/X$ for both $\dot{v}_k$ and $\hat{\mu}_k$, avoid linear models of $y_k = \hat{Y}_k/\hat{Y}$; as long as an 'optimal' model needs non-linear transformation of $\hat{Y}_k$ to y_k, the back-transformed $\hat{\mu}_k$ for θ_k will no longer be 'optimal', such that $\phi = 0$ would not minimise the TMSE(ϕ)
- Needs 'exchangeability' for $(\hat{\mu}_k, \hat{\phi})$ in $\mathcal{D}_1$ to benefit $k \in \mathcal{D}_0$, which is reasonable if SRS of areas; separating take-somes from take-alls benefits transductive estimation
- MSE or interval estimation for $k \in \mathcal{D}_0$ possible based on exchangeability, either assumed or is approximately justified by the sampling design
- Synthetic estimation is a 'bold' alternative to EBLUP in SAE, because bias for model-outlier area is unlimited; transductive estimator provides extra robustness, when model-outliers to μ_k caused by *other* covariates than q_k (MNO-MINDS, 2025, Ch. 11)

TL for flash estimation

Generic setup for flash estimation (Zhang & Haug, 2025)

- available: $\mu(x, s)$ from s; of interest: y_j for $j \in R$
- $f(y_j \mid x_j)$ for $j \in R$ *known* to differ to $f(y_i \mid x_i)$ for $i \in s$



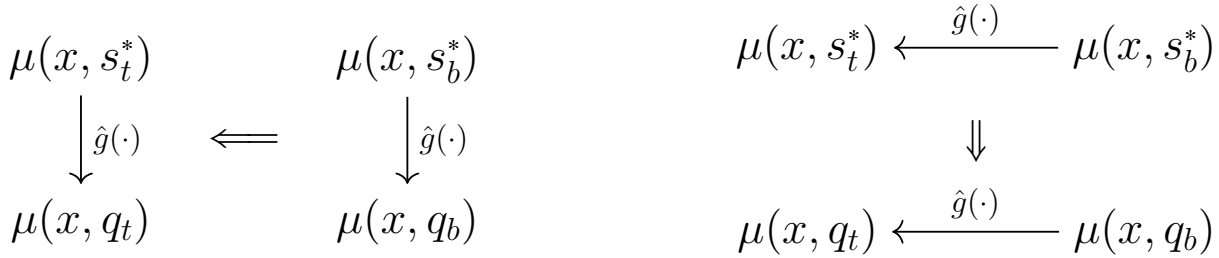
Augmented sample $s^* = s_t \cup R_{t-1}$ (left) or $s^* = s_t \cup R_{t-1} \cup R_{t'}$ (right)

Augmented learning: un-regularised loss, augmented s_t^*

$$\min_\beta \sum_{i \in s_t^*} \{x_i^\top \beta - \mathbb{I}(i \in s_t) y_i - \mathbb{I}(i \in r_t^*) \tilde{y}_i\}^2 \quad \text{given} \quad s_t^* = r_t^* \cup \underbrace{s_t \cup \cdots \cup s_t}_{\gamma^{-1} \text{ times}}$$

$$\Leftrightarrow \quad \min_\beta \left(\sum_{i \in s_t} (x_i^\top \beta - y_i)^2 + \gamma \sum_{j \in r_t^*} (x_i^\top \beta - \tilde{y}_j)^2 \right)$$

TL for flash estimation



Quasi transfer learning (QTL), scheme A (left), scheme B (right)

Zhang and Haug (2025) propose augmented learning and QTL:

- augmented learning $\mu(x, s_t^*)$ for flash estimation of the Norwegian Retail Turnover Index, within 2 (instead of 4) weeks and only based on stores that can report via headquarters within a week after each calendar month, i.e. s_t by purposive sampling
  NB. unacceptable results using $\mu(x, s_t)$ by supervised-learning
- apply QTL to squared prediction error of $\mu(x, s_t^*)$, where b is set by the latest $\tilde{y}$ available at t; resulting confidence interval has about the same width as the design-based confidence interval (based on existing larger sample), with somewhat conservative coverage

Since absence (or lack) of target labels is the characteristic challenge for flash estimation, how to organise source data is a conceptual issue beyond supervised learning (given any models).

Summary

By following assumption-lean predictive inference approaches, it is possible to maintain the rigour of predictive inference for official statistics, which is concerned with supervised learning given arbitrary ML models that are patently not 'true', whether the associated uncertainty is assessed in terms of bias, MSE or prediction interval.

Transfer learning offers a general assumption-lean non-Bayesian paradigm in settings where unbiased supervised learning is too noisy or impossible.

Brief mentions of future development

- Off-the-shelf tools for practitioners

- SRB-based predictive inference (quasi-randomisation): nonresponse imputation & out-of-sample prediction

- General conditions of pq-design and model training for unbiased out-of-sample prediction? Debiasing ML for Neyman-Fisher consistency generally?

- Beyond unconditional coverage of unit-level marginal prediction intervals, how to achieve suitable conditional average that varies with the features?

- Extend transfer learning applications & novel learning approaches, where supervised learning is inadequate

...

Bibliography

Barber, R.F., Candès, E.J., Ramdas, A. and Tibshirani, R.J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10:455-482.

Beaumont, J.-F. and Haziza, D. (2022). Statistical inference from finite population samples: A critical review of frequentist and Bayesian approaches. *The Canadian Journal of Statistics*, 50:1186-1212.

Berger, Y. G. (1998). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 67:209-226.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, 18: 105-110.

Boistard, H., Lopuhaä, H. P. and Ruiz-Gazan, A. (2017). Functional central limit theorems for single-stage sampling designs. *The Annals of Statistics*, 45:1728-1758.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32:190-205.

Dagdoug, M., Goga C., and Haziza, D. (2023). Model-Assisted Estimation Through Random Forests in Finite Population Sampling. *Journal of the American Statistical Association*, 118:1234-1251

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fay, R.E. and Herriot, R.A. (1979) Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 85:398-409.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh and London.

Fuller, W. A. (2009). *Sampling Statistics*. Wiley.

Geisser, S. (1993). *Predictive Inference*. Chapman & Hall.

Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations I. *The Annals of Mathematical Statistics*, 36:1707-1722.

Gu, T., Han, Y. and Duan, R. (2025). Robust angle-based transfer learning in high dimensions. *Journal of the Royal Statistical Society Series B*, 87:723-745.

Bibliography

Hansen, M. (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, 2:180-190.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78:776-793.

Han, Q. and Wellner, J.A. (2021). Complex sampling designs: Uniform limit theorems and applications. *The Annals of Statistics*, 49:459-485.

Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Hartley, H. O. and Ross, A. (1954). Unbiased Ratio Estimators. *Nature*, August 7, pp. 270-271.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy*, 5:361-374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35:1491-1523.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663-685.

James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 1:361-379.

Kalton, G. (2002). Models in practice of survey sampling. *Journal of Official Statistics*, 18:129-154.

Lee, D., Zhang, L.-C. and Chen, S. (2023). Robust quasi-randomization-based estimation with ensemble learning for missing data. *Scandinavian Journal of Statistics*, 50:1263-1278.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113:1094-1111.

Li, S., Cai, T.T. and Li, H. (2020). Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *Journal of the Royal Statistical Society Series B*, 84:149-173.

Lumley, T. (2024). `survey`: analysis of complex survey samples. R package version 4.4.

Bibliography

Mickey, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54:594-612.

MNO-MINDS (2025). *Methodologies and open source tools for integrating MNO and non-MNO data sources, Deliverable 3.2*. ESSnet MNO-MINDS. `https://cros.ec.europa.eu/system/files/2025-09/WP3_D3.2.pdf`

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society, pp. 558-625.*

Pan, S.J. and Yang, Q. (2009). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345-1359.

Ranalli, M.G. (2025). Machine learning methods for estimation in Official Statistics. *Journal of Official Statistics*, 41:912-920.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 37:81-91.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

Rao, J. N. K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31:117-138.

Rao, J. N. K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26:240-256.

Rao, J. N. K. and Wu, C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society B*, 72:533-544.

Sanguiao-Sande, L. and Zhang, L.-C. (2021). Design-Unbiased Statistical Learning in Survey Sampling. *Sankhya A*, Centenary Issue in Honour of C. R. Rao, 83:714-744.

Shafer G. and Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9:371-421.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer, New York.

Bibliography

Smith, T.M.F. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society, Series A*, 146:394-403.

Smith, T. M. F. (1994). Sample surveys 1975–1990; an age of reconciliation? (with discussion). *International Statistical Review*, 62:5-34.

Tano, B. and Zhang, L.-C. (2025). Combining Travel Survey and Mobile Network Operator data to produce disaggregated trip statistics. *To be submitted*.

Tibshirani, R.J., Barber, R.F., Candes, E. and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.

Tuoto, T. and Zhang, L.-C, (2025). Assessing uncertainty when integrating Mobile Network Operator data into official statistics with some transfer learning methods. *ITACOSM 2025, Bologna*.

Valiant, L. (2013). *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books, New York.

Valliant, R., Dorfman, R. M., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Vovk, V., Gammerman, A., and Shafer, G. (2022). *Algorithmic learning in a random world, 2nd Edition*. Springer.

Wieczorek, J. (2924). Design-based conformal prediction, *Survey Methodology*, 49:443-473.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185-193.

Zhang (2025). Some transfer learning methods for official statistics. *SIS Conference 2025, Genoa*.

Zhang, L.-C. and Haug, J.K. (2025) Turnover Flash Estimation by Purposive Sampling and Debit Card Transactions. *Journal of Official Statistics*, 41:382-403.

Zhang, L.-C. and Sanguaio-Sande, L. (2025). Design-based prediction intervals. *To be submitted*.

Zhang, L.-C., Sanguiao-Sande, L. and Lee, D. (2025). Design-based predictive inference. *Journal of Official Statistics*, 41:404-432.

Zhang, L.-C. and Haraldsen, G. (2022). Secure Big Data Collection and Processing: Framework, Means and Opportunities. *Journal of the Royal Statistical Society Series A*, 185:1541-1559.