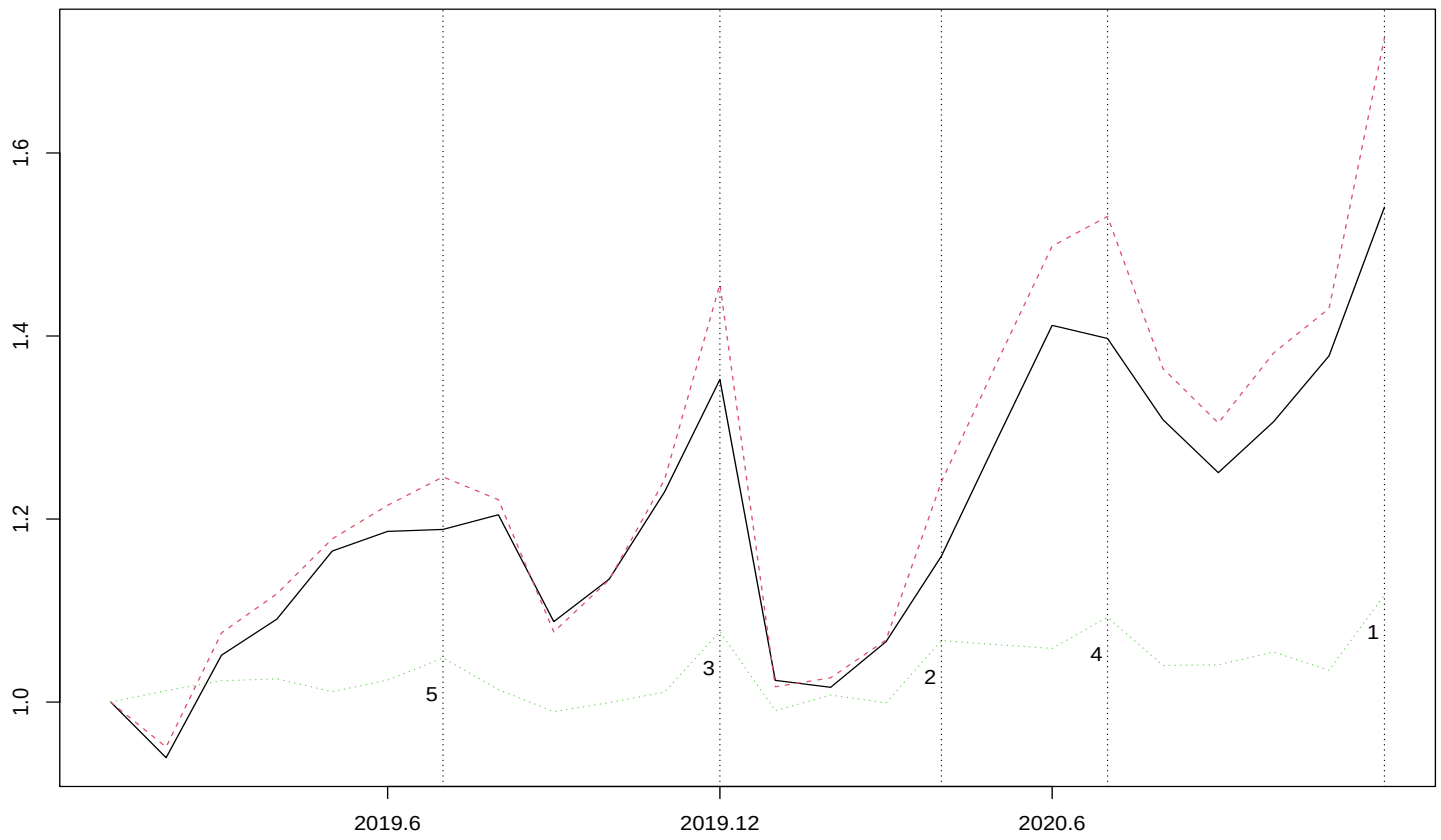
Turnover flash estimation using debit card payment data

Li-Chun Zhang[1,2] **and** ***Jens Kristoffer Haug***[1]

[1]***Statistisk sentralbyrå, Norway***
[2]***University of Southampton (L.Zhang@soton.ac.uk)***

Some background



Retail Turnover Index (solid) and a **transaction index (dashed)**
in Norway over 24 months in 2019 and 2020.
Five largest fluctuations of month-on-month ratio between them
(dotted, faint) marked by vertical lines (1, ..., 5).

Some background

$t + 30$: Retail Turnover Index (RTI) for NACE45-47
For each month t, the sample consists of $s_t \cup r_t$ with

- *take-all* units s_t, self-representing
- *take-some* units r_t, representing $R_t = U_t \setminus s_t$

Aim: flash estimation by $t + 15$***, survey only*** s_t

- reduced survey burden and processing resource
- requires appropriate modelling and learning, since s_t is not 'representative' of whole population U_t

Available data:

- $\{y_{ti} : i \in s_t\}$ **survey turnover**, monthly
- $\{y^*_{ti} : i \in U^*_t\}$ past **VAT turnover**, months delay
- $\{z_{ti} : i \in U^B_t\}$ **debit card payment** total, daily
  NB. generally, $U_t \neq U^*_t \neq U^B_t$, $y_{ti} \neq y^*_{ti} \neq z_{ti}$

A generic setup for flash estimation

$y =$ target outcome and $x =$ associated features
Denote by $\mu(x, s)$ a predictor for any unit with features x, obtained from

$$\{(y_i, x_i) : i \in s\}$$

where s is the *training* sample of observations. Denote by R a *target* set of units with known x_j, $\forall j \in R$, and

$$s \cap R = \emptyset,$$

for which the predicted y-values are of interest.
However, it is known that $\mu(x, s)$ is biased for $\{y_j : j \in R\}$, because y_j for $j \in R$ and y_i for $i \in s$ do not have the same distribution conditional on x_j and x_i.

Fundamental challenge:
lack of observations from target domain,
e.g. $R_t = U_t \setminus s_t$ for RTI

Learning approaches, given model $\mu(x)$

SIMPLISTIC LEARNING: only based on take-all sample s_t

$$\{(x_i, y_i) : i \in s_t\}$$

AUGMENTED LEARNING: augment s_t by additional data

$$s_t^* = s_t \cup r_t^*, \quad \{(x_i, y_i) : i \in s_t\}, \quad \{(x_j, y_j^*) : j \in r_t^*\}$$

NB. Given any constant $\gamma > 0$, *augmented loss*

$$L(s \cup r^*; \gamma) = \sum_{i \in s} \{\mu(x_i) - y_i\}^2 + \gamma \sum_{j \in r^*} \{\mu(x_j) - y_j^*\}^2$$

where r^* denotes a set of units that are similar to or even overlap with those in R, but y_j^* is a proxy to target y_j also if $j \in r^* \cap R$. Now,

$$\begin{aligned}
\hat{\beta} &\overset{\mu(x)=x^\top\beta}{=} (\textstyle\sum_{i \in s} x_i x_i^\top + \gamma \sum_{j \in r^*} x_j x_j^\top)^{-1} (\sum_{i \in s} x_i y_i + \gamma \sum_{j \in r^*} x_j y_j^*) \\
&= \arg\min_\beta L(s \cup r^*; \gamma) \;=\; \arg\min_\beta L(s^*) \\
L(s^*) &= \sum_{i \in s^*} \{\mu(x_i) - y_i\}^2 \quad \text{and} \quad s^* = r^* \cup s \cup \cdots \cup s
\end{aligned}$$

where s is duplicated γ^{-1} times in s^* (if practically possible)

Learning approaches, given model $\mu(x)$

y_{ti} = month-t turnover, x_{ti} = VAT turnovers, transactions
Augmented sample

$$s_t^* = s_t \cup r_t^*$$

Setting-I:

$$r_t^* = r_{t'}, \quad r_t^* = r_{t-1} \quad \text{or} \quad r_t^* = r_{t'} \cup r_{t-1}$$

where $r_{t'}$ is r-sample (of take-some units) for the same month previous year, and r_{t-1} that of previous month
NB. y_j^* for $j \in r_t^*$ is a *past survey turnover* value

Setting-II:

$$r_t^* = R_{t'}^*, \quad r_t^* = R_{t-d}^* \quad \text{or} \quad r_t^* = R_{t'}^* \cup R_{t-d}^*$$

where $R_{t'}^*$ contains all non-take-all units with *VAT turnover* for same month previous year, and R_{t-d}^* for month $t-d$
NB. $d = 4$ in Norway to ensure y_j^* is available, $\forall j \in r_t^*$

NB. in either setting, many units in r_t^* still belong to R_t
Augmented learning can help if $E(y_{ti} \mid x_{ti})$ is similar to $E(y_{t^*i}^* \mid x_{t^*i})$ given sensible choice t^* or composition of r_t^*

Learning approaches, given model $\mu(x)$

Augmented learning using s_t and additional units

Setting-I			Additional units r_t^*		
t	t'	$t-1$	(a)	(b)	(c)
Sept'22	Sept'21	Aug'22	$r_{\text{Sept'21}}$	$r_{\text{Aug'22}}$	$r_{\text{Sept'21}} \cup r_{\text{Aug'22}}$
Aug'22	Aug'21	Jul'22	$r_{\text{Aug'21}}$	$r_{\text{Jul'22}}$	$r_{\text{Aug'21}} \cup r_{\text{Jul'22}}$
⋮					
Sept'21	Sept'20	Aug'21	$r_{\text{Sept'20}}$	$r_{\text{Aug'21}}$	$r_{\text{Sept'20}} \cup r_{\text{Aug'21}}$
Setting-II, $d=4$			Additional units r_t^*		
t	t'	$t-d$	(a)	(b)	(c)
Sept'22	Sept'21	May'22	$R^*_{\text{Sept'21}}$	$R^*_{\text{May'22}}$	$R^*_{\text{Sept'21}} \cup R^*_{\text{May'22}}$
Aug'22	Aug'21	Apr'22	$R^*_{\text{Aug'21}}$	$R^*_{\text{Apr'22}}$	$R^*_{\text{Aug'21}} \cup R^*_{\text{Apr'22}}$
⋮					
Sept'21	Sept'20	May'21	$R^*_{\text{Sept'20}}$	$R^*_{\text{May'21}}$	$R^*_{\text{Sept'20}} \cup R^*_{\text{May'21}}$

Learning approaches, given model $\mu(x)$

Transfer learning: *target* model $\mu(x;\beta)$ and sample s; *source* model $\mu(x;\hat{\theta})$ estimated from a similar population; obtain $\hat{\beta}$ based on $L(\beta;s)$ but subjected to $\|\hat{\theta}-\beta\|_2 \leq \epsilon$

Quasi transfer learning: target $\mu(x,q_t)$ for domain q_t

$$\begin{array}{ccc} \mu(x,s_t^*) & \xleftarrow{\hat{g}(\cdot)} & \mu(x,s_b^*) \\ \Downarrow & & \\ \mu(x,q_t) & \xleftarrow{\hat{g}(\cdot)} & \mu(x,q_b) \end{array} \qquad \text{given } b<t$$

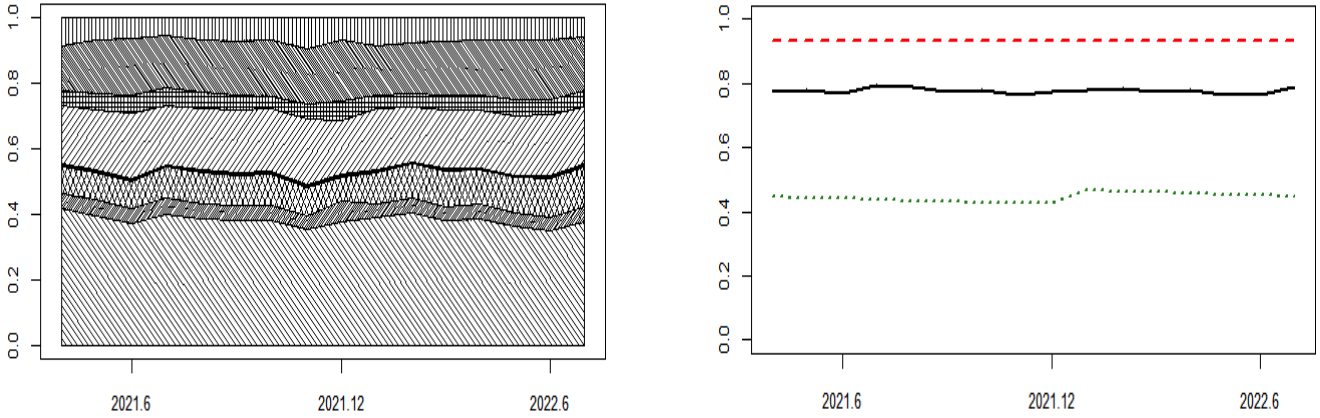
$$E\{\mu(x,s_t^*)\} = g\big(x,\mu(x,s_b^*)\big)$$

$t=$ Sept'22	Setting-I, $b=$ Aug'22	Setting-II, $b=$ May'22
Target q_t	$s_{\text{Sept'22}} \cup r_{\text{Sept'22}}$	$s_{\text{Sept'22}} \cup R_{\text{Sept'22}}$
Source:target q_b	$s_{\text{Aug'22}} \cup r_{\text{Aug'22}}$	$s_{\text{May'22}} \cup R_{\text{May'22}}$
Source s_t^*	$s_{\text{Sept'22}} \cup r_{\text{Sept'21}} \cup r_{\text{Aug'22}}$	$s_{\text{Sept'22}} \cup R^*_{\text{Sept'21}} \cup R^*_{\text{May'22}}$
Source:source s_b^*	$s_{\text{Aug'22}} \cup r_{\text{Aug'21}} \cup r_{\text{Jul'22}}$	$s_{\text{May'22}} \cup R^*_{\text{May'21}} \cup R^*_{\text{Jan'22}}$

Model learning illustrated

Major subdivisions of NACE47 - Retail sales

471	Non-specialised stores
472	Food, beverages and tobacco in specialised stores
473	Automotive fuel in specialised stores
474	Information & communication equipment in specialised stores
475	Other household equipment in specialised stores
476	Cultural and recreation goods in specialised stores
477	Other goods in specialised stores
478	Via stalls and markets
479	Not in stores, stalls or markets



Left, domain turnover shares. Right, s_t turnover share (solid), population proportion (dotted), sample proportion (dashed).

Model learning illustrated

Setting-I: relative average MSE in 2022

	Sample for learning				
NACE471	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-1}$	$s_t \cup r_{t'} \cup r_{t-1}$	$s_t \cup r_t$
Linear regression	1	1.00	1.00	1.00	1.00
Random forest	1	0.99	0.65	0.65	0.24
NACE475	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-1}$	$s_t \cup r_{t'} \cup r_{t-1}$	$s_t \cup r_t$
Linear regression	1	1.02	0.93	0.94	0.86
Random forest	1	1.04	0.86	0.92	0.21
NACE477	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-1}$	$s_t \cup r_{t'} \cup r_{t-1}$	$s_t \cup r_t$
Linear regression	1	0.92	0.83	0.77	0.78
Random forest	1	0.54	0.52	0.47	0.28

Note: MSE = $\sum_{i \in r_t} \{y_{ti} - \mu(x_{ti})\}^2 / |r_t|$

Residual for hypothetical learning on $s_t \cup r_t$

Augmented learning better than simplistic learning

Random forest usually better than linear regression

Model learning illustrated

Setting-I: average SME ($\times 10^2$) in 2022

	Sample for learning				
NACE 471	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-1}$	$s_t \cup r_{t'} \cup r_{t-1}$	$s_t \cup r_t$
Linear reg.	310	306	309	305	308
Random forest	268	214	167	157	58
NACE 475	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-1}$	$s_t \cup r_{t'} \cup r_{t-1}$	$s_t \cup r_t$
Linear reg.	201	192	113	117	132
Random forest	260	108	220	161	16
NACE 477	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-1}$	$s_t \cup r_{t'} \cup r_{t-1}$	$s_t \cup r_t$
Linear reg.	161	263	133	188	107
Random forest	118	56	37	37	8

Note: SME = $\{|r_t|^{-1} \sum_{i \in r_t} \mu(x_{ti}) - y_{ti}\}^2$

SME more relevant than MSE for Official Statistics

Residual for hypothetical learning on $s_t \cup r_t$

Random forest usually better than linear regression

Model learning illustrated

Setting-II: average SME ($\times 10^2$) in 2022

	Sample for learning				
NACE 471	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-4}$	$s_t \cup r_{t'} \cup r_{t-4}$	$s_t \cup r_t$
Linear reg.	310	298	307	297	309
Random forest	254	200	226	188	235
NACE 475	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-4}$	$s_t \cup r_{t'} \cup r_{t-4}$	$s_t \cup r_t$
Linear reg.	201	156	206	168	155
Random forest	266	110	188	93	123
NACE 477	s_t	$s_t \cup r_{t'}$	$s_t \cup r_{t-4}$	$s_t \cup r_{t'} \cup r_{t-4}$	$s_t \cup r_t$
Linear reg.	161	264	203	246	116
Random forest	109	38	38	18	34

Note: using VAT instead of survey turnover in r_t^*

Random forest better than linear regression

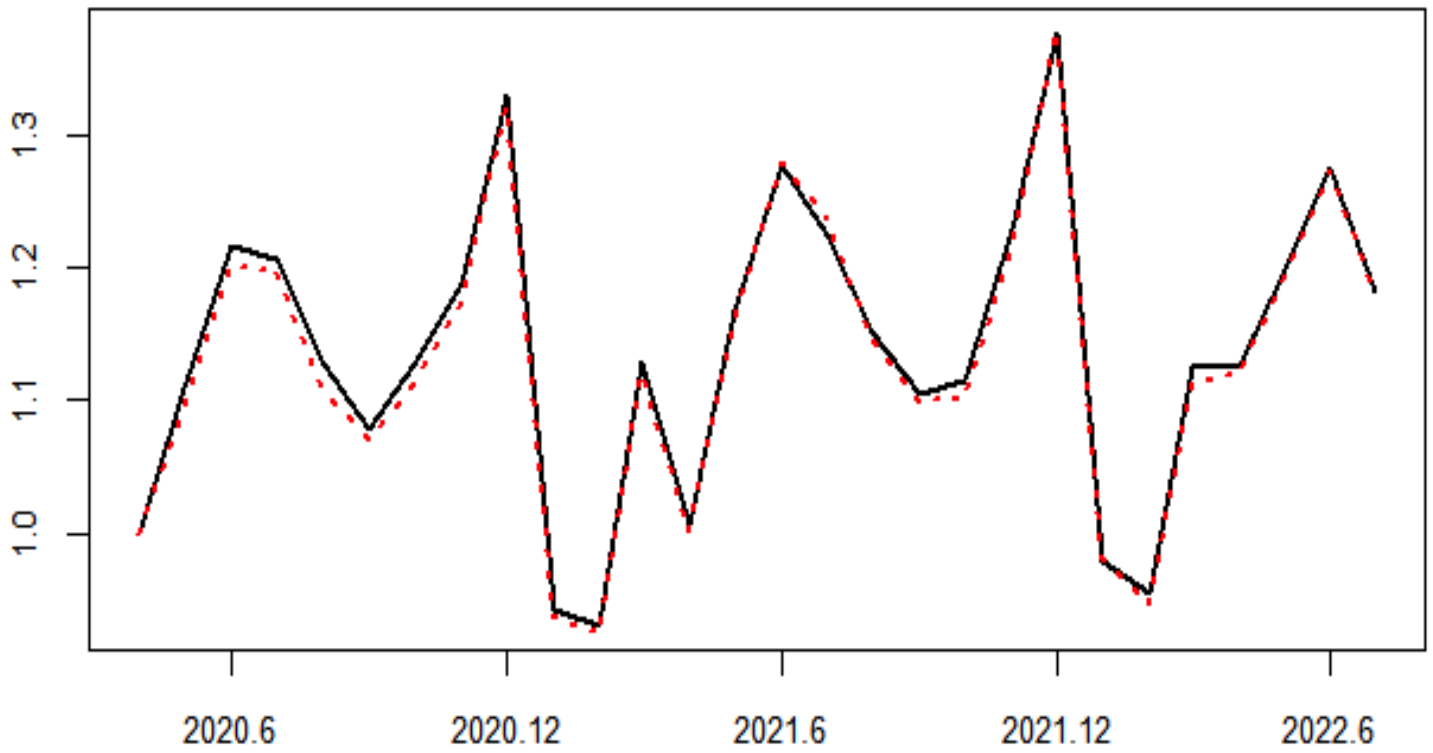
Augmented $s_t^* = s_t \cup r_t^*$ better than simplistic s_t

Can accept $\{y_{ti} : i \in s_t\} \cup \{y_{tj}^* : j \in r_t^*\}$ for $\{y_{ti} : i \in s_t \cup r_t\}$, in particular, $r_t^* = r_{t'} \cup r_{t-4}$ seems a robust choice

Flash RTI for NACE47

RTI (solid) for NACE47 over periods of 2021 - 2022
flash RTI: simplistic (dotted), **augmented (dashed)**
hypothetical learning from r_t (cross)
NB. production implementation currently at SSB

Retrospective validation



RTI (solid) and **VAT index (dotted)** in Norway over 2020-2022

NB. Distinguish **modelling** to **learning** (details omitted here)

NB. One should be able to tell if the chosen model and learning approach have worked satisfactorily, till as recently as, say, d months ago, had VAT turnover been the target measure. Details of methods omitted here.

Real-time uncertainty assessment

Retrospectively observable index I_t^* based on

survey $\{y_{ti} : i \in s_t\}$ and VAT $\{y_{tj}^* : j \in U_t^* \setminus s_t\}$

Flash index $\hat{I}_t$ (by $t + 15$) based on

observed $\{y_{ti} : i \in s_t\}$ and predicted $\{\hat{y}_{tj} : j \in U_t \setminus s_t\}$

Aim: ***prediction interval*** *of* I_t *at the time of* $\hat{I}_t$

Empirical calibrated $[\hat{I}_t - \delta_t,\ \hat{I}_t + \delta_t]$ where

$$\delta_t = \arg\min_{\substack{\delta \\ \delta>0}} \Big(\sum_{b=t-d}^{t-D} \mathbb{I}(\hat{I}_b - \delta \le I_b^* \le \hat{I}_b + \delta) = 1 - \frac{1}{D-d+1} \Big)$$

i.e. δ_t is the minimum positive value of δ such that the interval $[\hat{I}_b - \delta, \hat{I}_b + \delta]$ achieves the specified coverage over the most recent time window $b \in \{t-d, ..., t-D\}$
E.g. if $D - d + 1 = 12$, empirical coverage $= 11/12 = 91.7\%$

Real-time uncertainty assessment

Error-prediction by **Quasi Transfer Learning**:

$$[\hat{Y}_t^c - \alpha_t \hat{\sigma}_t,\ \hat{Y}_t^c + \alpha_t \hat{\sigma}_t] \quad \text{for} \quad Y_t^c = \textstyle\sum_{j \notin s_t} y_{tj}^*$$

with empirically calibrated α_t given $\hat{\sigma}_t^2$, which is given by

$e_{tj} = \mu(x_{tj}, s_t^*) - y_{tj}^*$ by augmented learning

$\sigma_t^2 \coloneqq V\Big(\sum_{j \notin s_t} e_{tj}\Big) = \sum_{j \notin s_t} E(e_{tj}^2)$ of independent errors

$\hat{e}^2 = \eta(x, \tilde{s}_t^c)$ error prediction model η

$\Rightarrow \quad \hat{\sigma}_t^2 = \sum_{j \notin s_t} \eta(x_{tj}, \tilde{s}_t^c)$ learned from $\tilde{s}_t^c$

Illustration: $r_t^* = \tilde{s}_t^c = R_{t'}^* \cup R_{t-d}^*$, $d = 4$

t	t'	$t-d$	t''	$t'-d$	$(t-d)'$	$(t-d)-d$
Sept '22	Sept '21	May '22	Sept '20	May '21	May '21	Jan '22
⋮						
May '22	May '21	Jan '22	May '20	Jan '21	Jan '21	Sept '21

Real-time uncertainty assessment

Two remarks:

- *error prediction* is a distinct learning task to *outcome prediction* because, without any target observations (in R_t), one cannot pretend the adopted model, e.g. $\mu(x, s_t^*)$, would yield unbiased prediction of the target outcome (e.g. turnover) and derive the associated uncertainty as a by-product of the outcome model;
- to improve the efficiency of error prediction one should utilise observed unit-level prediction errors in the past just like one uses observed past outcomes at the unit level for outcome prediction.

Inference for NACE47

VAT total (solid), prediction interval by empirical method (dotted),
error prediction (dashed), CI by survey sampling (long-dashed).
NB. relative half-length 4.7% by error prediction,
7.8% by empirical method, and 4.6% by survey sampling
NB. nominal level 91.7% for prediction interval, 95% for CI;
coverage 100% for prediction interval, 91.7% for CI

Some final remarks

Flash RTI for NACE47 is possible by augmented learning, which halves the current dissemination time lag, reduces the response burden and processing cost, without compromising the accuracy of RTI by traditional survey.

To achieve official statistics quality, we have developed methods of retrospective validation of model / learning, as well as real-time prediction interval estimation.

In addition to debit card data used here, other types of transaction are also of interest, such as e-invoices and business-to-business bank transfers. Despite distinct challenges of access and processing, they complement each other in coverage and content, becoming more useful in combination with each other. This is an area that requires strategic development of knowledge, experience and capacity at NSOs. A coordinated program across all the business statistics would be more impactful.

Some final remarks

Provided greater access to relevant and timely non-survey big data, improving the timeliness of economic indicators while reducing the response burden and processing cost becomes an ever more urgent matter. Survey data may still be necessary to ensure the relevance and accuracy of official statistics, as it is in the case of RTI. Combining appropriate purposive samples with *novel* modelling and learning approaches may be considered in practice.

Research of machine learning for official statistics may be less concerned about which model to use than how to organise the data outside the target domain but can be relevant (for training any given models), such as

augmented learning, quasi transfer learning.

For official statistics repeated over time and geography, various forms of transfer learning seem a large topic for future research and application.