Metodutmaningar för officiell statistik

Officiell statistik i en digital värld – att gå från data till statistik

- SCB och Örebro universitets sommarskola i statistik 2021

Dr Anders Holmberg
Chief Methodologist-General Manager
Australian Bureau of Statistics

Tack till kollegor på ABS för original till några illustrationer

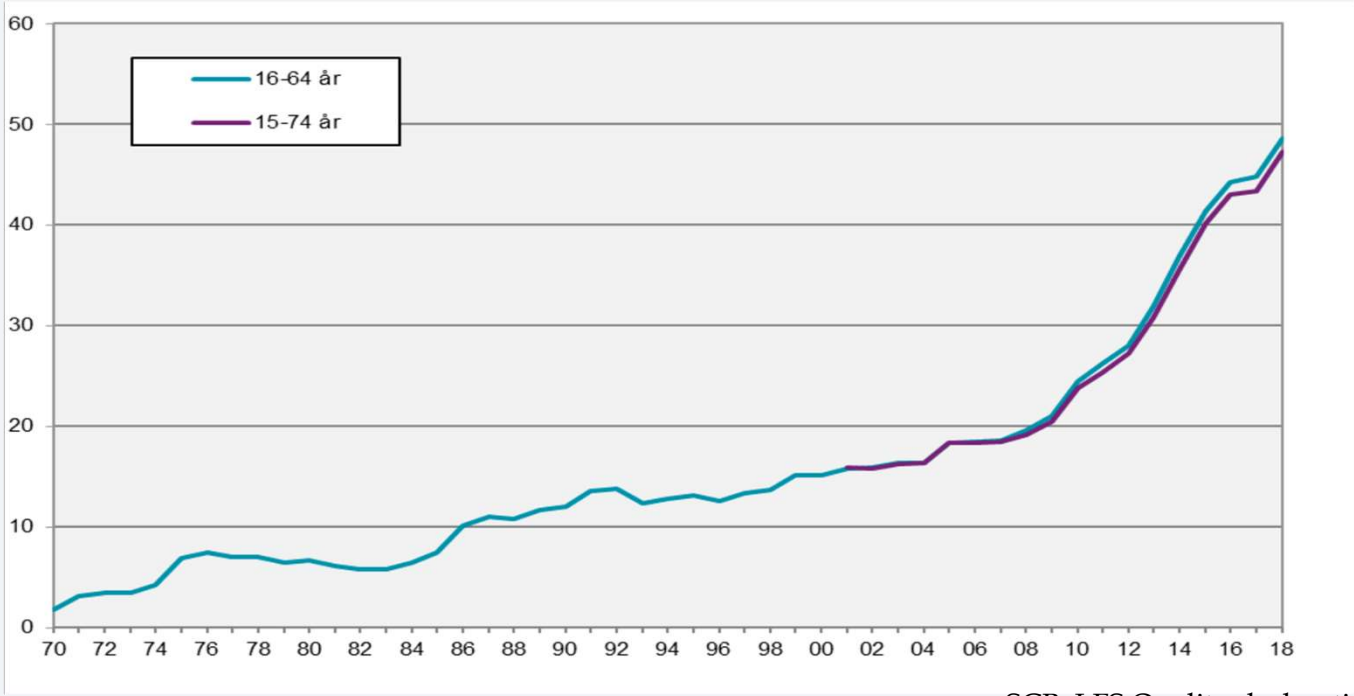
Australian Bureau of Statistics
Informing Australia's important decisions

Disposition

- Gamla problem i ny tappning, eller?
- Tankegods och Trender
  - Statistik i en reformerad fabrik?
  - Ny arbetsmodell med plattformar för ökad datatillgänglighet
  - Eller både och?
- Några Exempel

Senaste 20 åren. Bortfallet är ett gissel, vad kan göras?

Bortfall i AKU, 1970-2018



SCB: LFS Quality declaration version 1 2019-04-09

Nya datakällor med data som kan bli statistik, (kanske)

Australian Bureau of Statistics

Surveys and Censuses

Administrative Collections e.g. STP

Product Scanners

Smart Meters

Transactions data e.g. bank data

Imagery Systems

Telematics Devices

Personal Devices

Environmental Monitors

Logistics Systems

Social Media Services

Tankegods och trender: Miljö

- Gemensamma system och dataförvaltning i det allmänna
  - ….och i övriga samhället
- ‘Ekosystem‘ för data och för datarelaterade tjänster (statistisk analys och forskning)
- Nya teknologiska framsteg och lösningar, molntjänster, parallel computing
  - ..grafdatabaser, maskininlärning, privacy-preserving techniques, resampling, mikrosimulering, optimering, avancerade geospatiala lösningar

Tankegods och trender: Krav

- Ny statistik med ökad betoning på aktualitet och detaljrikedom men också...
- Dynamiskt användarorienterat utbud av statistik, **RELEVANS**
- Öka tillgången, dela och utnyttja existerande datakällor. Forskning och analys, automatisering och förenkling för de som lämnar uppgifter
- Skydd av data och skydd av personlig integritet

Tankegods och trender: Möjligheter

- Applicera en kunskapssyn på livscykeln för data.
- Mer av data integration och kopplingar av olika datakällor
- Ny tillämpning för statistisk metod som datavetenskap. Omvärdera gamla designer i ny kontext? Nya ramverk? Bortfallsundersökningar?
- Mer modellanvändning men krav på transparens och kommunikation.

En spännande tid med möjligheter och utmaningar

- Kan vi utnyttja alla källor med de metoder, den teknologi och den lagstiftning som finns?
- Finns det kapacitet, kunskap och vilja?
- Här följer några exempel på hur flera datakällor kan användas!

Att kombinera primära (survey) och sekundära källor (register eller andra alternativa källor)

Förbättra skattningar

Förbättra aktualiteten på indikatorer

Nya analysupptäckter

Konfrontera och justera data från alt. källor

Massimputering

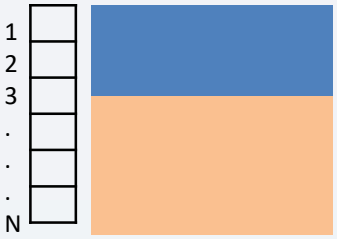
Syntetiserade data

Annat?

Lösningar med multipla datakällor är inget nytt

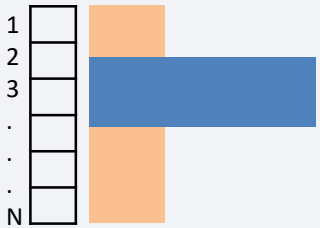
- Data från olika källor kan kombineras på manga sätt, (se t.ex ESSnet Komuso)
- Detta är praxis hos de flesta statistikproducenter.

Complementary, non-overlapping micro datasets

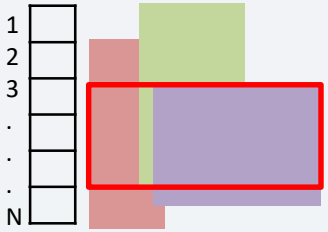


e.g. survey of largest units, register data for the rest

Overlapping micro datasets

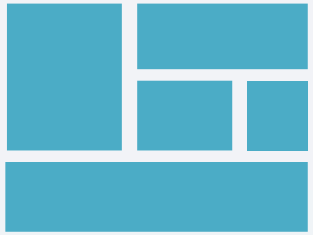


e.g. complete register and complementary sample survey data



e.g. record linkage to create a new data asset

Combining aggregates

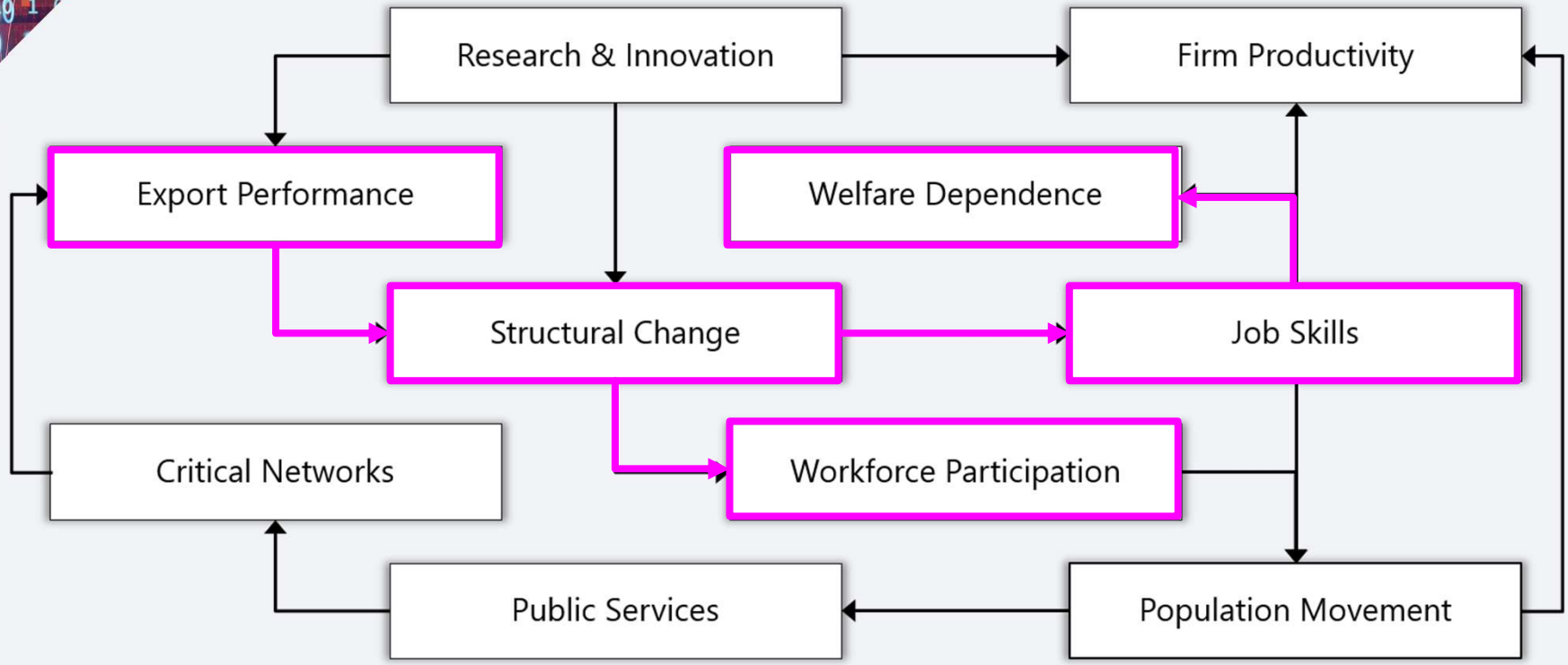


e.g. account compilation

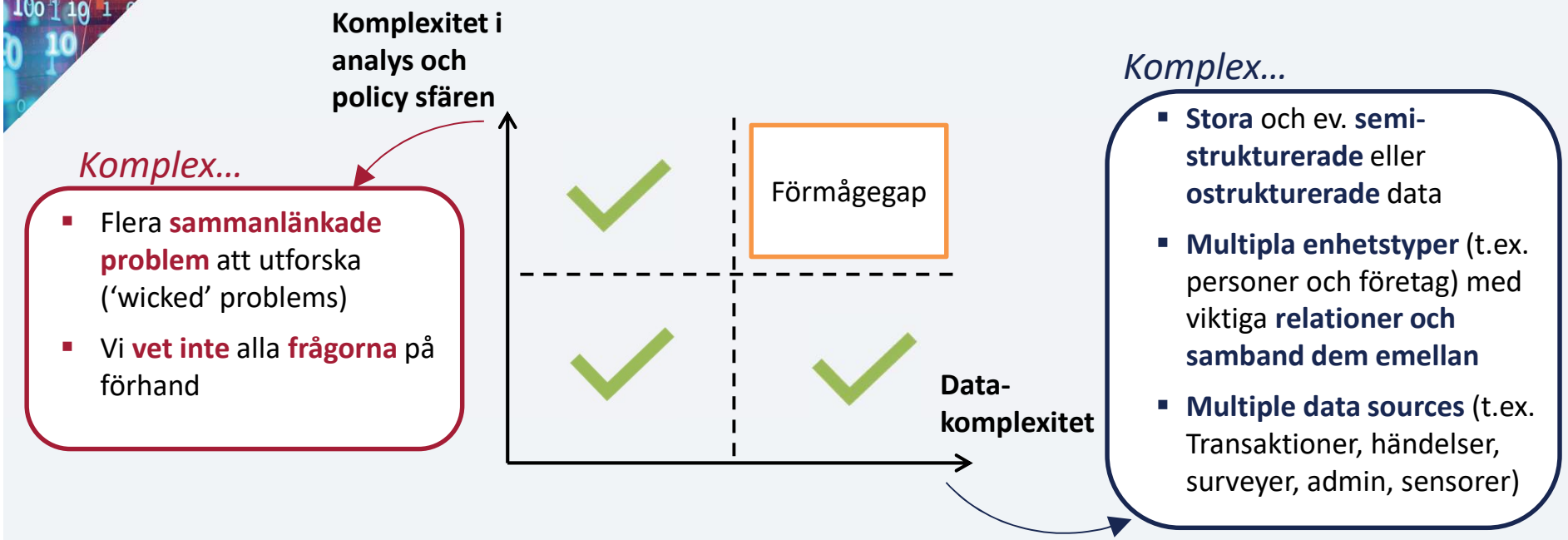
Exempel på informationsbehov

Export Performance	What are the characteristics of successful Australian exporters?	Merchandise exports data
Structural Change	In what industries, occupations and locations are jobs being created?	Business register information Business activity statements
Workforce Participation	What kinds of workers hold multiple jobs, and in what industries are these?	Job advertisements
Job Skills	Is higher education meeting the demand for skills in the new economy?	Tax records Student records
Welfare Dependence	What factors are associated with long term welfare dependence?	Social services payments

Exempel på ett komplext problem och datasammansättning



Exististerande förmåga att koppla samman data för att möta efterfrågan och nya möjligheter är otillräcklig



Att kombinera primära (survey) och sekundära källor (register eller andra alternativa källor)

Nya analyser och upptäckter

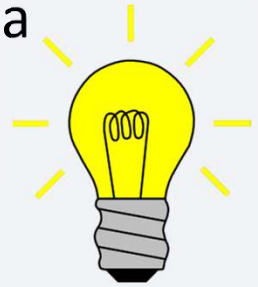
Massimputering

Förbättra aktualiteten på indikatorer

Syntetisera data

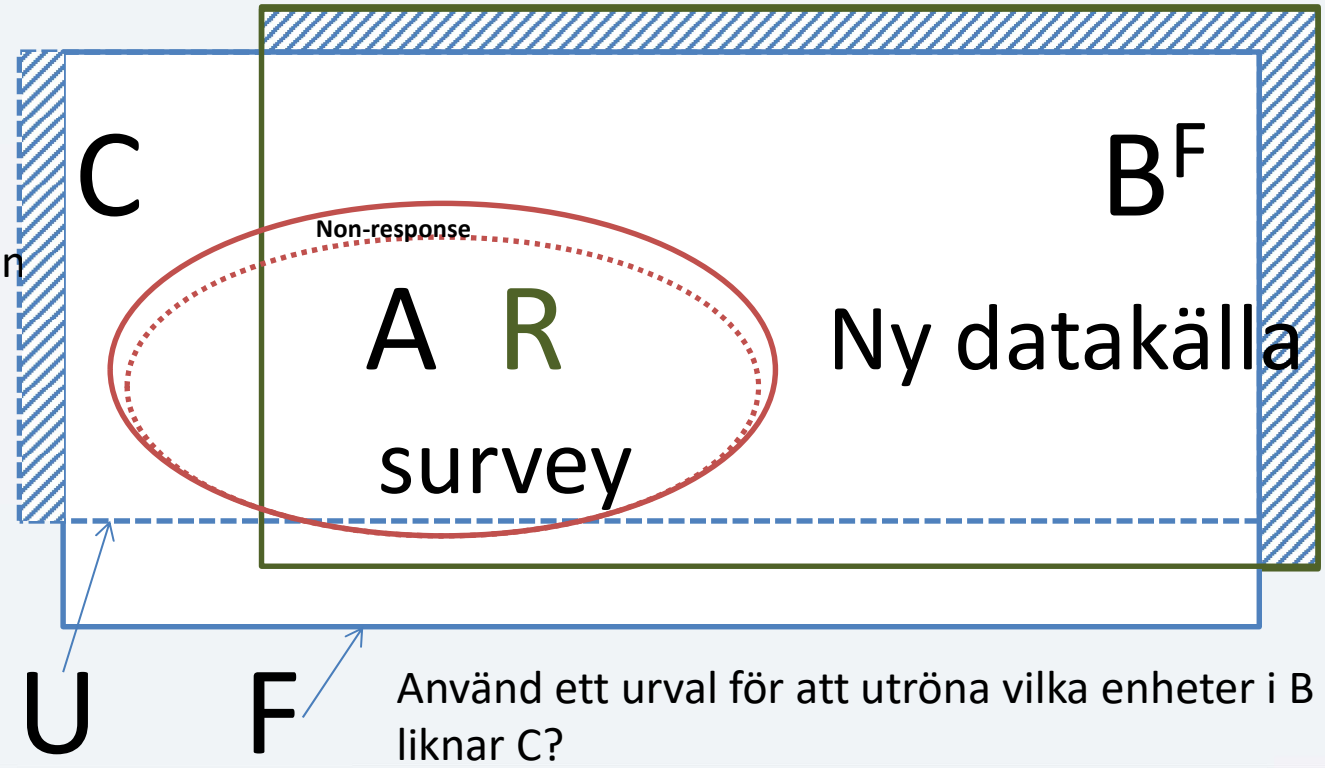
Konfrontera och justera data från alt. källor

Förbättra skattningar



Annat?

Dataset: Population, Nya (stora), Urval och ramar



Exkluderar populationen U som inte tacks av ramen F

Länkning av B till F gör att vi exkluderar (big) data enheter som inte är i ramen

Använd ett urval för att utröna vilka enheter i B som liknar C?

Datastrukturer (Tam and Holmberg, 2020)

	Source	Response variable, Y	Auxiliary variable, X	Representativeness?
TYPE 1	New data source, B	A	A	No
	Additional source, A	NA	A	No
TYPE 2	New data source, B	A	A	No
	Probability sample survey source, A	NA	A	Yes
TYPE 3	New data source, B	A	A	No
	Probability sample survey source, A	A	A	Yes
TYPE 4	New data source, B	NA	A	No
	Probability sample survey source, A	A	A	Yes

Note: A denotes available, NA denotes not available

Datastrukturer (cont'd)

- Exempel
  - Typ I – B: On line panel med mål och hjälpvariabler + A: kvoturval med endast hjälpvariabler.
  - Typ II – B: On line panel med mål och hjälpvariabler + sample survey covariate data only
  - Typ III – B: On line panel och A: Sannolikhetsurvalsbaserad survey båda med mål och hjälpvariabler
  - Typ IV – Satellitdata med enbart våglängdsdata och en jordbrukssurvey med mål och våglängdsdata

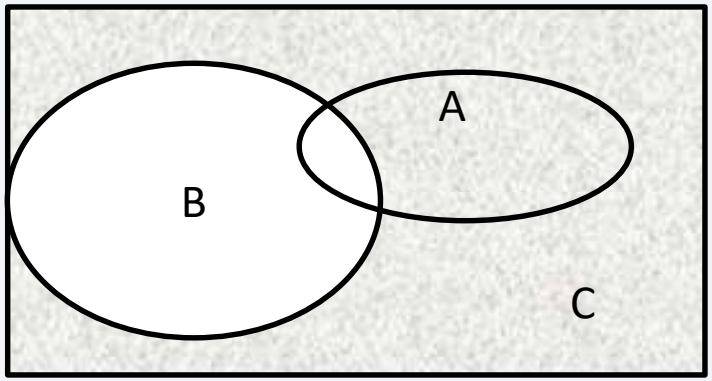
The set up

- Let $U = \{1, \ldots, N\}$ represent the labels of a finite population, $B \subset U$ represent those in the new data source, $A \subset U$ is a probability sample, with $|B| = n_B$, $|A| = n_A$, and the new data source indicator, $\delta_i = 1$, if $i \in B$, or 0 other wise.
- Assume that the study variable, y_i, is observed without error (in this talk) for $i \in B$, and the covariate, x_i is without error for every $i \in U$.
- Let $\bar{y}_B = \frac{\sum_{i \in B} y_i}{n_B}$, and $\bar{y}_A = \frac{\sum_{i \in A} d_i y_i}{n_A}$, where d_i is the sample weight. We are interested to estimate $\mathrm{T} = \sum_{i \in U} y_i$, or equivalently $\overline{\mathrm{T}} = \sum_{i \in U} y_i / N$.

Skattningsmetoder för Datastruktur Typ 3 (Kim & Tam, 2020)

- MAR antagande ej nödvändig men alla enheter I U måste ha en positiv chans att väljas.
- Undersökningsvariabeln i B och A observeras utan fel
- En Post-stratifiering estimator för totalen.

$\hat{T}_{PS} \triangleq \sum_{i\in B} y_i + N_C \frac{\sum_{i\in A}(1-\delta_i)d_i y_i}{\sum_{i\in A}(1-\delta_i)d_i}$ är approximativt unbiased och konsistent



Regression Data Integration Estimator (RegDI)

- Låt $\nu_i = (1 - \delta_i,\ \delta_i,\ \delta_i y_i)^T$ och $\nu = \sum_{i \in U} \nu_i$.
- Definiera $\hat{T}_{RegDI} \triangleq \sum_{i \in A} w_i y_i$, där $w_i = d_i \nu^T (\sum_{i \in A} d_i \nu_i \nu_i^T)^{-1} \nu_i$.
- Då blir $\hat{T}_{RegDI} = \hat{T}_{PS}$.
- RegDI formuleringen är användbar då den kan:
  - Inlemma mer hjälpinformation
  - Hantera kopior av enheterna i det nya datasetet B
  - Hantera bortfall i A
  - Hantera mätfel i B eller A.
  - Hantera länkningsfel mellan B och A

Exempel på RegDI varianter

- Mer hjälpinformation
  - Använd $v_i = (1 - \delta_i,\ \delta_i,\ \delta_i y_i, \delta_i x_i)^T$ eller $(1 - \delta_i,\ \delta_i,\ \delta_i y_i, x_i)^T$ om x_i observeras för $i \in B$; eller om x_i observeras för $i \in U$, beräkna sedan nya vikter w_i

- Hantering av kopior av enheterna i det nya datasetet B
  - Ändra δ_i i v_i från ett-noll till antalet gånger enhet *i* i A matchas med B och beräkna nya w_i

Egenskaper för RegDI estimatorn

- B = 2015-16 Ag Census, med svarsandel på 85%
- A = 2014-15 Ag Survey med svarsandel på 78%
- Problem
  - Undertäckning in B
  - Bortfallsfel i A
- Trots problemen ovan så är RegDI skattningarna 8 till 12 ggr mer effektiva än estimatorn i enbart urvalsundersökningen

Bias, Variance and Mean Squared Error of Selected Agricultural Commodities at the Australian level

Variable	Estimator from	Bias ($x10^3$)	Var ($x10^9$)**	MSE ($x10^9$)
DAIRY	REACS only (A)	0.00	6.19	6.19
	Agricultural Census only (B)*	-362.45	0	131.37
	RDI using (A) and (B)	0.00	0.43	0.43
BEEF	REACS only (A)	0.00	85.00	85.00
	Agricultural Census only (B)*	-2,389.53	0	5,709.86
	RDI using (A) and (B)	0.00	6.79	6.79
WHEAT	REACS only (A)	0.00	171.29	171.29
	Agricultural Census only (B)*	-2,043.52	0	4,176.00
	RDI using (A) and (B)	0.00	20.83	20.83

Notes (1) * Estimated by the difference between the total from B and the published ABS estimate from the Agriculture Census adjusted for non-response.

(2) ** Bootstrap estimates from 100 bootstrap samples.

Urval av referenser för de olika datastrukturer

- Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* **115**, 2011-2021
- Kim, J.K. (2018). Unpublished survey data integration lectures delivered to the Australian Bureau of Statistics.
- Kim, J.K. and Tam, S.M. (2020). Data integration by combining big data and survey sample data for finite population inference. International Statistical Review. **88** doi:10.1111/insr.12434
- Kim, J.K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review* **87**, 177-191
- Kott, P., and Chang, T. (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse, *Journal of the American Statistical Association* **97**,1265-1275
- Rivers, D. (2007). Sampling for web surveys. In *Proceeding of Section on Survey Research Methods*. American Statistical Association.
- Tam, S.M. and Holmberg, A. (2020). New data sources for official statistics – a game changer for survey statisticians? The Statistician **81**, 21-35.
- Yang, S. and Ding, P. (2018). Combining multiple observational data sources to estimate causal effects, *arXiv preprint arXiv:1801.00802* .
- Yang, S. and Kim, J.K. (2019). Nearest neighbour imputation for general parameter estimation in survey sampling. The Econometrics of complex survey data **39,** 209-234.
- Yang, S. and Kim, J.K. (2018). Integration of survey data and big observational data for finite population inference using mass imputation. https://arxiv.org/abs/1807.02817
- Yang, S., Kim, J.K. and Y. Hwang (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. Survey Methodology **47** 1 29-89
- Zhang, L.C. (2019). On valid descriptive inference from non-probability sample. The Statistical Theory and Related Fields 3, 103-113.

Att kombinera primära (survey) och sekundära källor (register eller andra alternativa källor)

Nya analyser och upptäckter

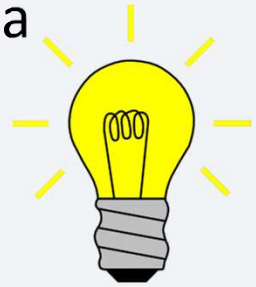
Massimputering

Förbättra aktualiteten på indikatorer

Syntetisera data

Konfrontera och justera data från alt. källor

Förbättra skattningar



Annat?

Exemplet efter lunch använder grafdatabaser och nätverk för att utforska multipla datakällor

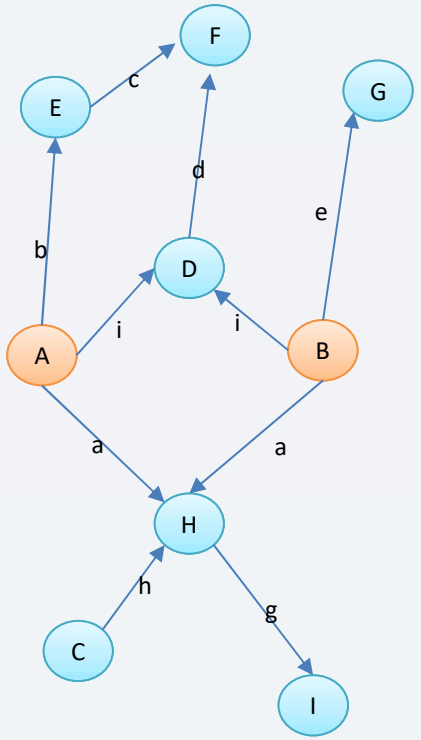
- En graf är ett matematiskt objekt

  G= (V, E, L)

  – *V:* set av noder (hörn – entiteter)

  – *E:* set av kanter (kopplingar – relationer)

  – *L:* set av etiketter (på kanterna eller noderna)
- Kan ha en riktning mellan noder eller vara utan
- Ofta använt för att beskriva modeller för sociala nätverk
- En naturlig representation av enheter och deras inbördes relationer

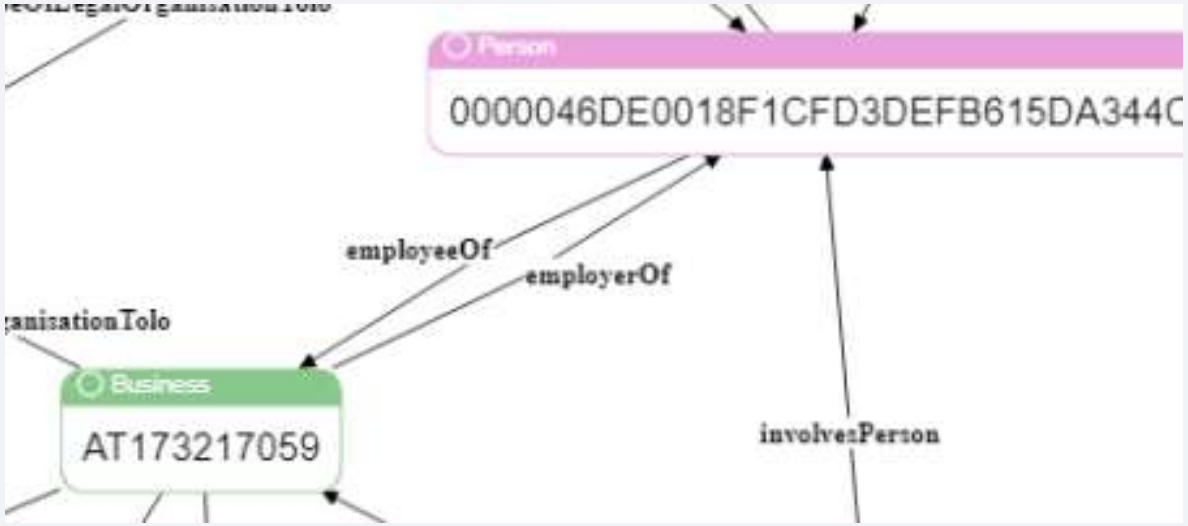


Graf data

- Lagrade och anropade som triplar i en grafdatabas:

  Subject: Predicate: Object

  Business: employerOf: Person



En blick runt hörnet

Kort sikt

- Vem vill tvätta data?
- Kunskapssyn på nya datakällor (utvärdera och övervaka antaganden)
- Designa för att autentisera data och uppskatta / justera för selektions- och mätbias
- Se till att teknologi och metoder stöder lösningar som är efterfrågestyrd data innovation och bevisa värdet.
- Hantera kvalitetskrav från mikrodataanvändare separat från statistiska kvalitetskrav på output
- Koppla samman data på lämpliga aggregeringsnivåer. (Mer geodata)
- Hantera data med etiska regler och skydda data!
- Nya insamlingsmetoder och urvalsdesigner

En blick runt nästa hörn

Längre sikt

- Etablera stabila populationsramar till statistik
- Semantisk märkning av data och dataset
- Använd (och neutralisera) tidsdimensionen vid analys av flera källor


Australian
Bureau of
Statistics
Tack så mycket!
Frågor?