

Using Non-Survey Big Data to Improve the Quality of the Household Budget Survey

29TH NORDIC STATISTICAL MEETING – REYKJAVIK – AUGUST 2022

MARIUS RUNNINGEN LARSSON, STATISTICS NORWAY – RIU@SSB.NO

LI-CHUN ZHANG, STATISTICS NORWAY

Background

The Household Budget Survey

- Measures the distribution of households' expenditure on goods and services
- Questionnaire and diary
- Results used amongst other things to
 - Measure distributional effects of tax changes
 - Consumer price index/National accounts
 - Research on e.g. distributional implications of tax evasion

Problems

Diary

- High response burden
- Prone to errors
- Resource heavy for the NSI
- Expensive
- Quality issues

Om vekten gjelder før eller etter fiskeviltet er rense/slaktet.
Hvis varen ikke er merket med vekt, må du om mulig veie varen selv eller anslå vekten.
Hvis det er vanskelig å gjøre anslag: Beskriv varen så godt det lar seg gjøre (antall, varemerke, produsent med videre.) Husk å få med alt fra melk, brød og mineralvann til alkoholholdige drikkevarer.

Egenproduserte matvarer

Her registrerer du også egenproduserte matvarer, det vil si matvarer som kommer fra eget jordbruk eller egen hage, ville bær og sopp som man plukker selv, fisk fra eget yrkes- eller sportsfiske, kjøtt fra jakt på vilt og liknende. For slike varer trenger du ikke oppgi pris.

Se for øvrig eksemplene nedenfor.



	Dato	Hva slags vare ble kjøpt? Beskriv varen.	Kryss av hvis varen er kjøpt		Mengde	Hva kostet varen	
			i utlandet	på internett		Kroner	Øre
Eks.1	2/1	Lammestek, fersk			2,3 kg	269	90
Eks.2	2/1	Grovbrød			800 g	26	50
Eks.3	5/1	Jordbær selvplukket			4 kg		
Eks.4	6/1	Laks, røkt			250 g	39	90
Eks.5	9/1	Skummet melk			1 liter	10	90
Eks.6	13/1	Rødvin, kartong (kjøpt i Sverige, svenske kroner)	X		3 liter	149	00
1		Frokost skinke			?	29	00
2		Smoothie			0,3	17	90
3		Epler			?	23	90
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							

Can we completely remove the diary component of the HBS?

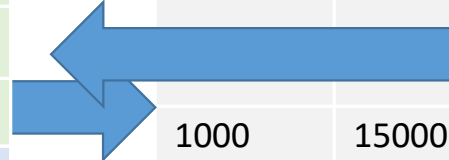
- Groceries accounts for approximately 60% of items in the diary
- Use data (receipts) from grocery stores (PoC 2018)
- Food and non-alcoholic beverages (COICOP 01)

Example

LI

Quantity	Price	COICOP	Location	Timestamp	Amount	...	ID
1	20,00	01.1.1.1	Store A	19:00:00	60,50	...	1
0,50	10,50	01.2.1.1	Store A	19:00:00	60,50	...	1
2	30,00	01.3.1.1	Store A	19:00:00	60,50	...	1
1	20,00	01.4.1.1	Store B	15:10:00	35,00	...	2
1	15,00	01.4.1.1	Store B	15:10:00	35,00	...	2
...

quantity	value
5000	100000
1000	15000
3000	52500
...	...



Sofiemyr

Apent 7 - 23 (7 - 23)
Telefon 66 82 10 00

Butikk 2705-3, Kaja
Kvittering 72248 2015-12-14 09:05

BAREPOSE	1.00
CASTELLO BLUE 150G	20.50
COCA COLA 1.5L	83.00
PANT	10.00
GILDE PINNEKJØTT	169.00
KREMFLØTE	16.90
KALROT	3.19
0.818 kg 3.90 kr/kg	
MELIS 500G	8.20
MUNKHOLM 0.33L	7.20
PANT	1.00
PIANO KAREMELLPUDDIN	37.80
REMULADE 165G	22.40
SANTA CLAUSTHALER	1.00
PANT	1.00
SMØR NORMALSALTET	25.50
TORO ERTESTUING	12.40
TORO KALRABISTAPPE	13.40
Totalt	433.00
Bank:	433.49
Herav	
Dagligvarer	433.49
Øvrige varer	0.00

income	...
1	...
1	...
1	...
3	...
...	...

Bax: 11576168-304590

MASTERCARD
*****1790-1
AID: A0000000041010
TVR: 0000008000
TSI: E800
Sted: 8067011

Receipts alone not sufficient

- No Personal Identifiable Information
- Need auxiliary information

KIWI 544 SØNSTERUDVEIEN
 ORG.NR. 937 846 231 MVA
 Foretaksregisteret
 SØNSTERUDVEIEN 28
 TLF: 66 82 18 65

14.12.15 09:22 Kasse:002
 Kvitt:7557 OperNr:111

COCA COLA 1,5LX4 FL	15%	83,00
+ Pant	0%	10,00
KARAMELLPUDDING 1L PIAN	15%	38,00
SANTA CLAUSTHALER 0,33L	15%	1,00
+ Pant	0%	1,00
ERTESTUING TORO	15%	12,40
KALRABISTAPPE 94G TORO	15%	13,40
MELIS 500G ELTORADO	15%	8,20
SMØR NORMAL 500G TINE	15%	25,50
KREMFLØTE 3DL TINE	15%	16,90
KALROT KG VASKET	15%	3,77
0,966kg x kr 3,90		
CASTELLO BLA 150G	15%	20,50
REMULADE EKTE 165G MILL	15%	22,40
PINNEKJØTT LAM GILDE	15%	169,00
BÆREPOSE KIWI	25%	0,99
SUBTOTAL		426,06
Sum 13 varer		426,06

BANK 426,06

Mva%	Grunnlag	Mva	Totalt
0,00	11,00	0,00	11,00
15,00	360,06	54,01	414,07
25,00	0,79	0,20	0,99
SUM	371,85	54,21	426,06

Bax: 12072155-295799
 BankAxept
 *****20704-1
 AID: D5780000021010
 TVR: 8000048000
 TSI: 6800
 14/12/2015 09:22
 Ref.: 237437 046220 IAI
 Resp.: 00
 KJØP
 NOK 426,06
 GODKJENT

REMA 1000
 REMA 1000 HOLMLIA
 Kovacs Dagligvare AS
 ORG NR. 999 250 386 MVA
 Foretaksregisteret
 TLF: 22 62 37 75

14.12.15 09:43 Kasse:001
 Kvitt:259826 OperNr:073

COCA-COLA 1,5L 4PK	15%	83,00
+ Pant	0%	10,00
PINNEKJØTT AV LAM	15%	169,00
MUNKHOLM ØL	15%	7,20
+ Pant	0%	1,00
BÆREPOSE	25%	0,99
KREMFLØTE	15%	16,90
KARAMELLPUDD.PIANO	15%	37,80
KALRABISTAPPE	15%	13,40
ERTESTUING GRØNN	15%	12,40
REMULADE	15%	22,40
BLA CASTELLO	15%	20,50
KALROT	15%	2,92
0,750kg x kr 3,90		
CLAUSTHALER SANTA	15%	1,00
+ Pant	0%	1,00
MELIS	15%	8,20
Sum 13 varer		407,71

BANK 407,71

Mva%	Grunnlag	Mva	Totalt
0,00	12,00	0,00	12,00
15,00	343,23	51,49	394,72
25,00	0,79	0,20	0,99
SUM	356,03	51,68	407,71

BAX: 473410-71402036
 14/12/2015 09:43
 2488534
 MasterCard
 *****1790
 AID:A0000000041010
 TVR:000008000
 TSI:E800
 REF:090 123400831882
 RESP:00 GODKJENT
 NOK= 407,71

Takk for handelen
 Facebook.com/REMA1000

coop extra
 Sofiemyr
 Apent 7 - 23 (7 - 23)
 Telefon 66 82 10 00

Butikk 2705-3, Kaja
 Kvittering 72248 2015-12-14 09:05

BÆREPOSE	1,00
CASTELLO BLUE 150G	20,50
COCA COLA 1.5L	83,00
PANT	10,00
GILDE PINNEKJØTT	169,00
KREMFLØTE	16,90
KALROT	3,19
0,818 kg 3,90 kr/kg	
MELIS 500G	8,20
MUNKHOLM 0.33L	7,20
PANT	1,00
PIANO KARAMELLPUDDIN	37,80
REMULADE 165G	22,40
SANTA CLAUSTHALER	1,00
PANT	1,00
SMØR NGRMALSALTET	25,50
TORO ERTESTUING	12,40
TORO KALRABISTAPPE	13,40
Totalt	433,00
Bank:	433,49

Herav
 Dagligvarer 433,49
 Øvrige varer 0,00

Bax: 11576168-304590
 MASTERCARD
 *****1790-1
 AID: A0000000041010
 TVR: 000008000
 TSI: E800
 Start: 2067011

Solution

Link receipts to card transactions

Method

Linkage Condition

Receipts - Card Transactions

JOIN CT with R on AMOUNT, LOCATION, TIMESTAMP

$R \bowtie_{\theta} CT$

$\theta = (amount_R = amount_{CT}) \wedge (location_R = location_{CT}) \wedge (timestamp_R = timestamp_{CT})$

Noise in the Key Variables

Amount

Location

Timestamp

Update Linkage Conditions

$$amount_R = amount_{CT}$$



$$amount_{R,f} = amount_{CT,f}$$
$$f \in \{charged, cost\}$$

$$location_R = location_{CT}$$



$$location_R^{DC} = location_{CT}$$

$$timestamp_R = timestamp_{CT}$$



$$timestamp_{R,end} < timestamp_{CT} + \tau_+$$
$$timestamp_{R,start} > timestamp_{CT} - \tau_-$$

Updated Linkage Condition

$$R \bowtie_{\theta(f, \tau_+, \tau_-)} CT$$

Initial conjunction rule

$$\theta = (amount_R = amount_{CT}) \wedge (location_R = location_{CT}) \wedge (timestamp_R = timestamp_{CT})$$

Updated conjunction rule

$$\begin{aligned} \theta(f, \tau_+, \tau_-) = & (amount_{R,f} = amount_{CT,f}) \wedge \\ & (timestamp_{R,end} < timestamp_{CT} + \tau_+ \mid timestamp_{R,start} > timestamp_{CT} - \tau_-) \wedge \\ & (location_R^{DC} = location_{CT}) \end{aligned}$$

Data

**Chain
#1**

Chain #2

Chain #3

**Debit card
transactions**

Table 4				
Table Key	$R^{processed}$	R^{medium}	R^{raw}	CT
Time	$timestamp_{R,start}$	$timestamp_{R,end}$	$timestamp_{R,end}$ $timestamp_{R,start}$	$timestamp_{CT}$
Amount	$amount_{R,cost}$	$amount_{R,cost}$	$amount_{R,charged}$ $amount_{R,cost}$	$amount_{CT,charged}$ $amount_{CT,cost}$
Location	$location_R^{DC}$	$location_R^{DC}$	$location_R^{DC}$	$location_{CT}$

- More than half a billion receipts
- More than 1.5 billion debit card transactions

Results

Results

	(τ_+, τ_-)	$R^{processed}$	R^{medium}	R^{raw}
MMR	(15, 75)	0.1324	0.0205	0.6795
	(30, 150)	0.0461	0.0206	0.0048
	(60, 300)	0.0306	0.0210	0.0038
PFL	(15, 75)	0.00315	0.00034	0.00006
	(30, 150)	0.00462	0.00143	0.00069
	(60, 300)	0.00694	0.00220	0.0017

Results summarized

- 98% of card transactions linked
- 74% of receipts linked
- 70% of receipts linked to households

Conclusion

Can we completely remove the diary component of the HBS?

Currently investigating methods for utilizing linked dataset (R-CT-AR-HH / Y) in the HBS

The goal is to reduce response burden in HBS 2026 (Groceries/ COICOP 01)

Thank you 😊

Marius Runningen Larsson – riu@ssb.no