Session 2.1

- The use of scanner data in Denmark- a case study

Data are validated in 3 steps

- First step: Simple checksystems in Sas (missing records, proper dates, number of records and format)
- Second step: SD is linked to the COICOP-classification based on the supermarkets' own classification and a searchword-process created in Excel and linked to Sas
- Third step: Further cleaning and aggregation of scanner data

Step 1: What is the output of validation-system

- Check for missing records
- Check for proper dates
- Check for number of records
- Check for format

EXERCISE 1 (5 minutes)

- What kind of check systems would you like to have for scanner data?

Step 2: Linking of Scanner data to ECOICOP

- The linking of GTIN codes to COICOP is done at a 6 digit level for each supermarket type of store. At the moment we use existing 8 digit level where possible

How is the linking done?

- First we have the nomenclature, which can be updated:

C6	C6_description	C6-description_extra
11110	Rice	Rice, pourridge of rice, ricedesserts
11121	Flower and grain	Wheat flower, wheat grains, other types of flower
11131	Ryebread	All kind of rye bread
11199	Residual group of bread and bakery products	

- Then we have the nomenclature linking (looked through once a year)

xxx_VNR	xxx_C6
xxxx	11110
xxxxx	11199

What do we do with the residual groups

C6	C6_description	Searchword1	Searchword2	New_C6	New_C6_description
11199	Residual group of bread and bakery products	%BASMATI%		11110	Rice
11199	Residual group of bread and bakery products	%HVEDEMEL%		11121	Flower and grain

How does this look in real life?

How do we monitor the new search words needed

C6	C6_description	aggr_turnover	turnover_c2_aggr	turnovershare of C2-level (PCT)	turnover_c4_aggr	turonvershare of C4-level (PCT)
22099	Restgrupper tobak	163.110	165.113	98,79	165.113	98,79
11399	Restgrupper fisk	3.202.789	1.447.052.237	0,22	52.538.203	6,1
11198	Restgrupper af bageri	8.491.246	1.447.052.237	0,59	184.288.644	4,61
11299	Restgrupper kød og fjerkræ	10.576.361	1.447.052.237	0,73	269.208.928	3,93

What isn't linked then?

Eancode	Product description	C6	C6 description	Ean-turnover	The eancodes' share of the residual group	The whole residual groups' share of the C2-level
1234567891011	LAKS	11399	Restgrupper fisk	299.3	8,06	0,22
1234567891011	MOERKSEJFILET	11399	Restgrupper fisk	183.9	4,95	0,22

And what do we do

- We use

C6	C6_description	Searchword1	Searchword2	New_C6	New_C6_description
11399	Residual group of fish	%MOERKSEJFILET%		11311	Cod and the like

Step 3: Aggregation

- This includes making amounts uniform
- It also includes cleaning other variables(product text, unit, amount…) at the lowest possible level

- For example GTINs with more than one amount are recalculated into numeraire
- Per default all PLUs with more than one product text, amount or unit are deleted here!

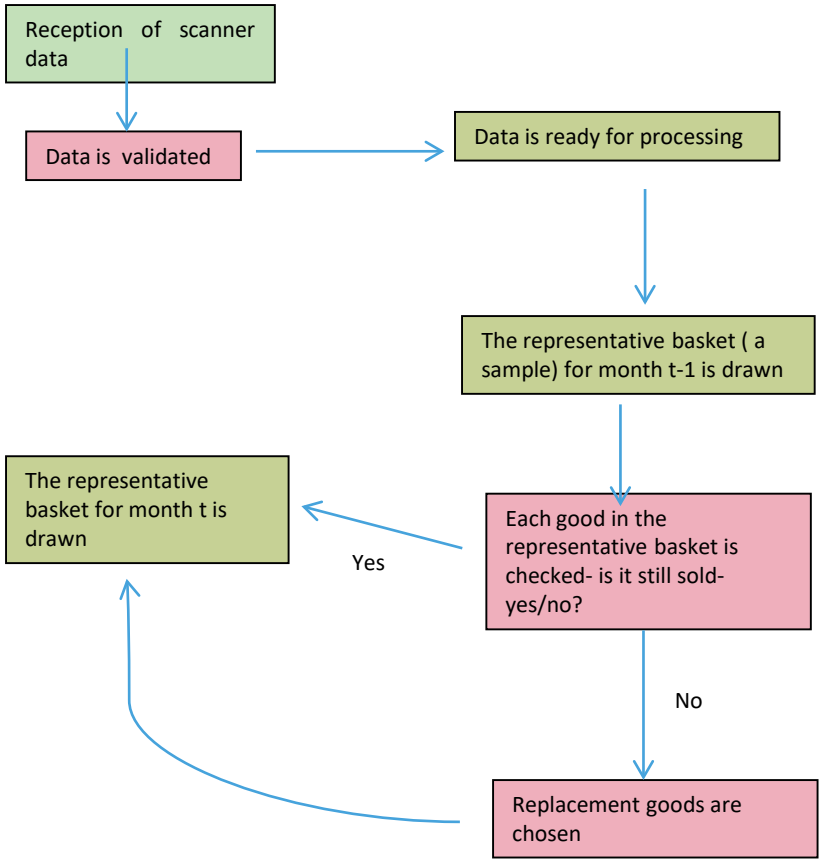
...

- It also includes removing prices, which are more than a factor 4 away from the median price for a given GTIN

- Now: The prices are aggregated into one price per GTIN per type of shop

- **The final step:**
- More weeks are aggregated together (including cleaning) to form the data for one month
- ***Application of filters:***
- Prices which have changed by more than a factor 10, since last month, are removed
- Prices which have fallen in both price and volume since last month are removed

CPI via scanner data



Selection criteria (1)

- Goods in the representative basket need to have been sold for all 12 months of 2011 (2015) and constitute the top 50 % of turnover.
- There are individual criteria at either too many or too few observations compared to the existing sample.

Selection criteria (2)

- Replacement goods need to have been part of the sample for 4 months
- They are then picked based on highest turnover and stability of turnover

System for maintenance of the representative basket

On the calculation of SD-indices

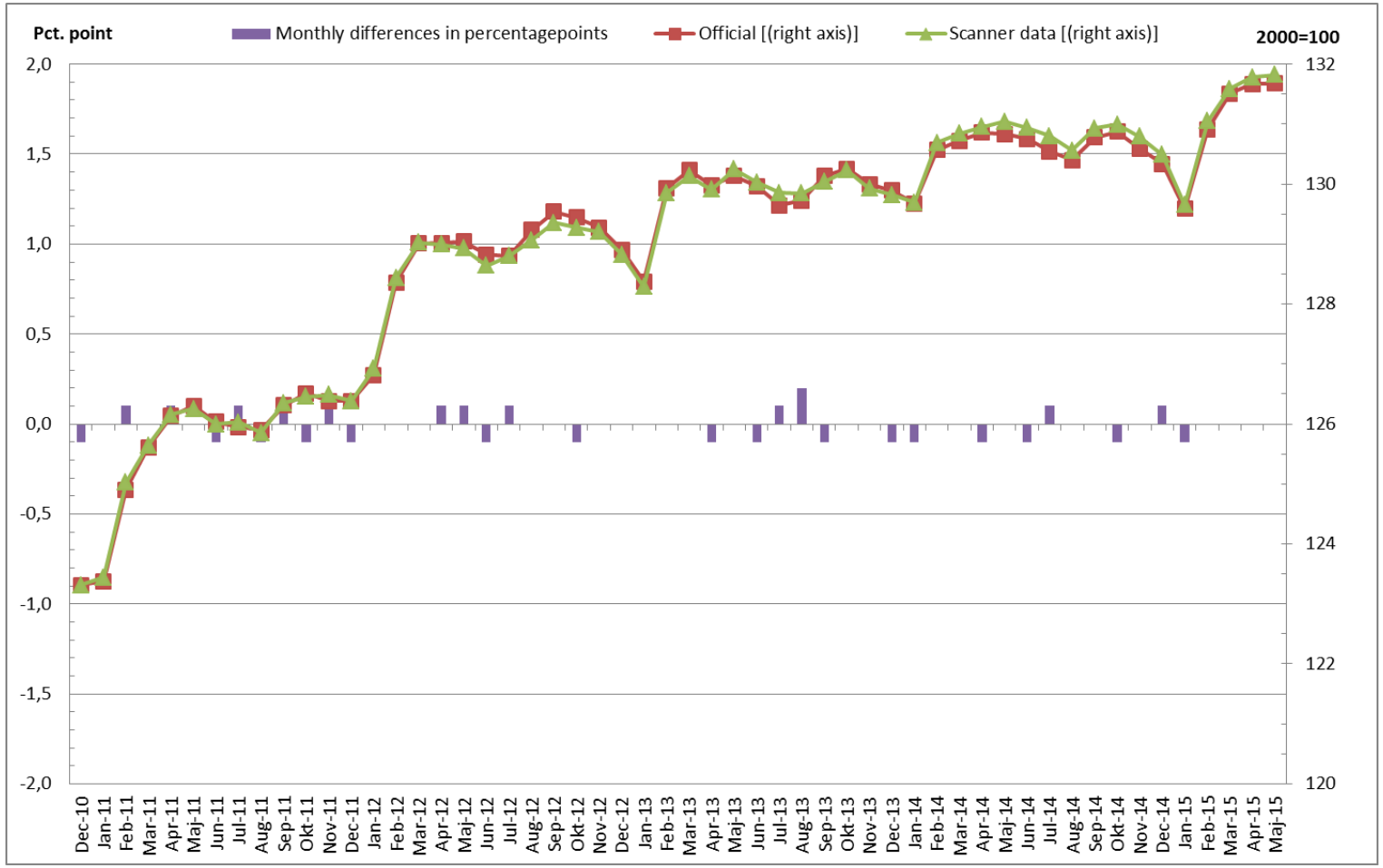
- Calculation is based on Laspeyres type indices (JEVONS)- done at all levels following the existing calculation methods

- February 2015: Commenced with the calculation of RYGEKS indices, so we are able to have a basis for a juxtaposition.

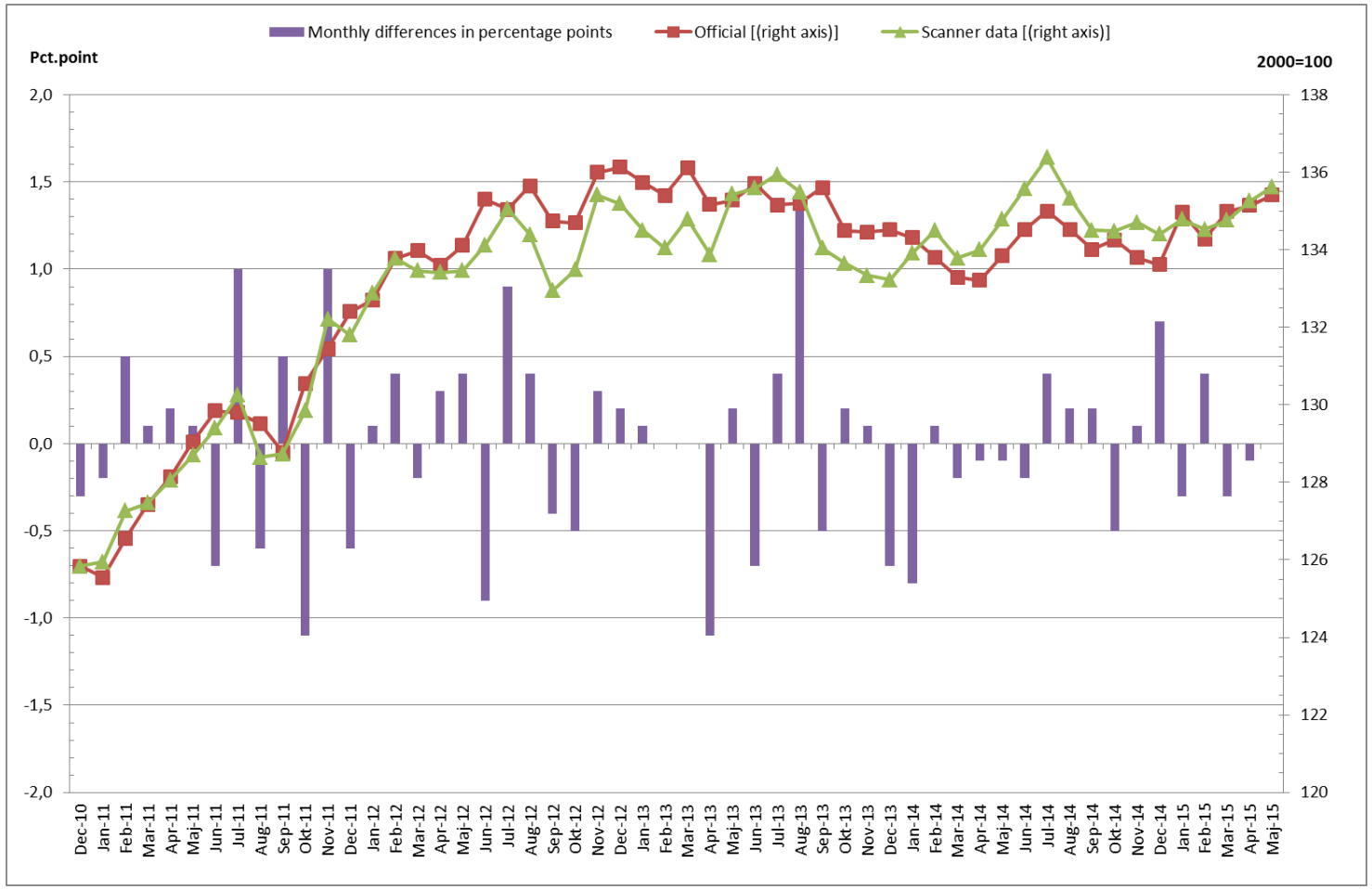
→ suggest small underestimation of price development, but rygeks method is not perfect

Documentation?

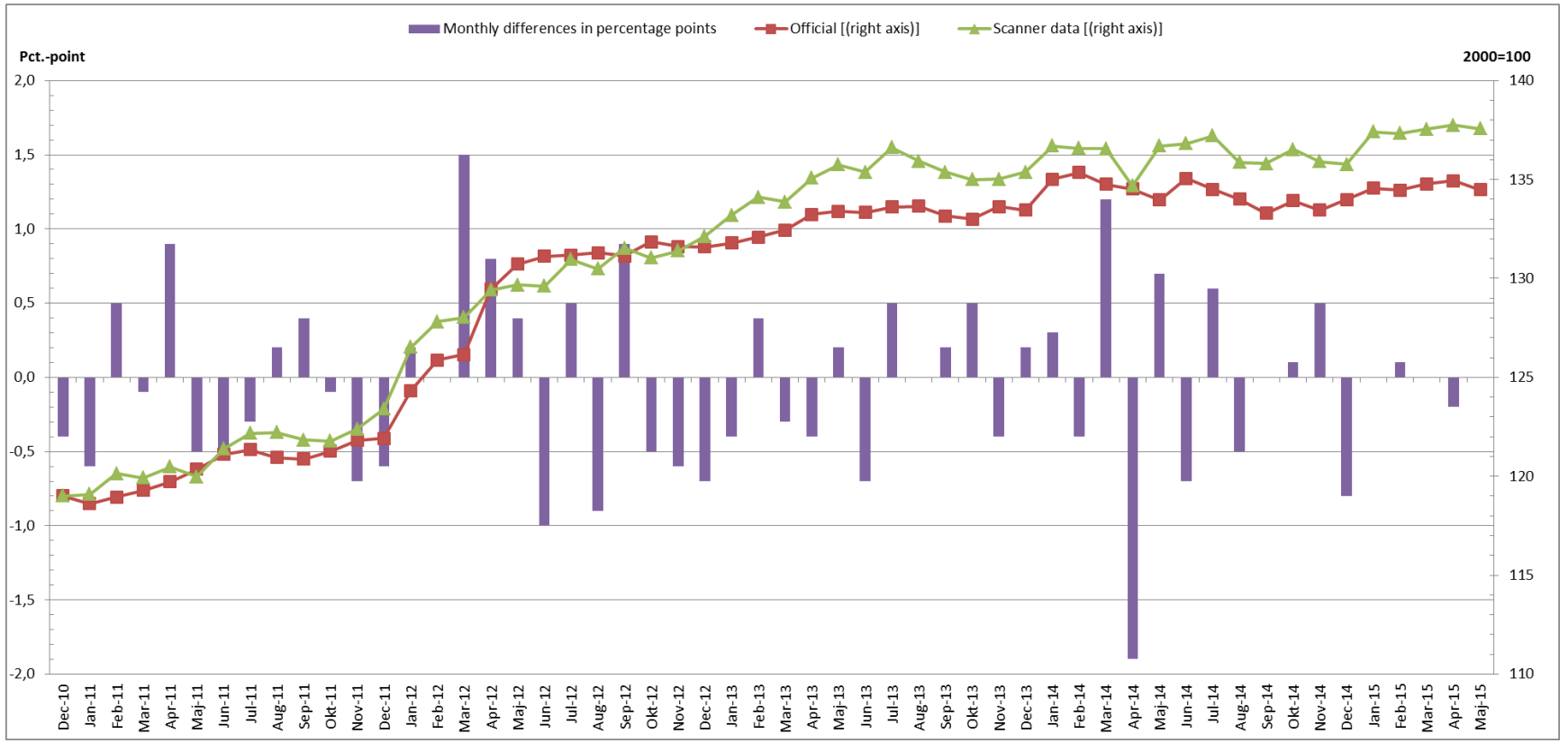
CPI


Pct. point
Monthly differences in percentagepoints
Official [(right axis)]
Scanner data [(right axis)]
2000=100
2,0
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
-2,0
132
130
128
126
124
122
120
Dec-10
Jan-11
Feb-11
Mar-11
Apr-11
Maj-11
Jun-11
Jul-11
Aug-11
Sep-11
Okt-11
Nov-11
Dec-11
Jan-12
Feb-12
Mar-12
Apr-12
Maj-12
Jun-12
Jul-12
Aug-12
Sep-12
Okt-12
Nov-12
Dec-12
Jan-13
Feb-13
Mar-13
Apr-13
Maj-13
Jun-13
Jul-13
Aug-13
Sep-13
Okt-13
Nov-13
Dec-13
Jan-14
Feb-14
Mar-14
Apr-14
Maj-14
Jun-14
Jul-14
Aug-14
Sep-14
Okt-14
Nov-14
Dec-14
Jan-15
Feb-15
Mar-15
Apr-15
Maj-15

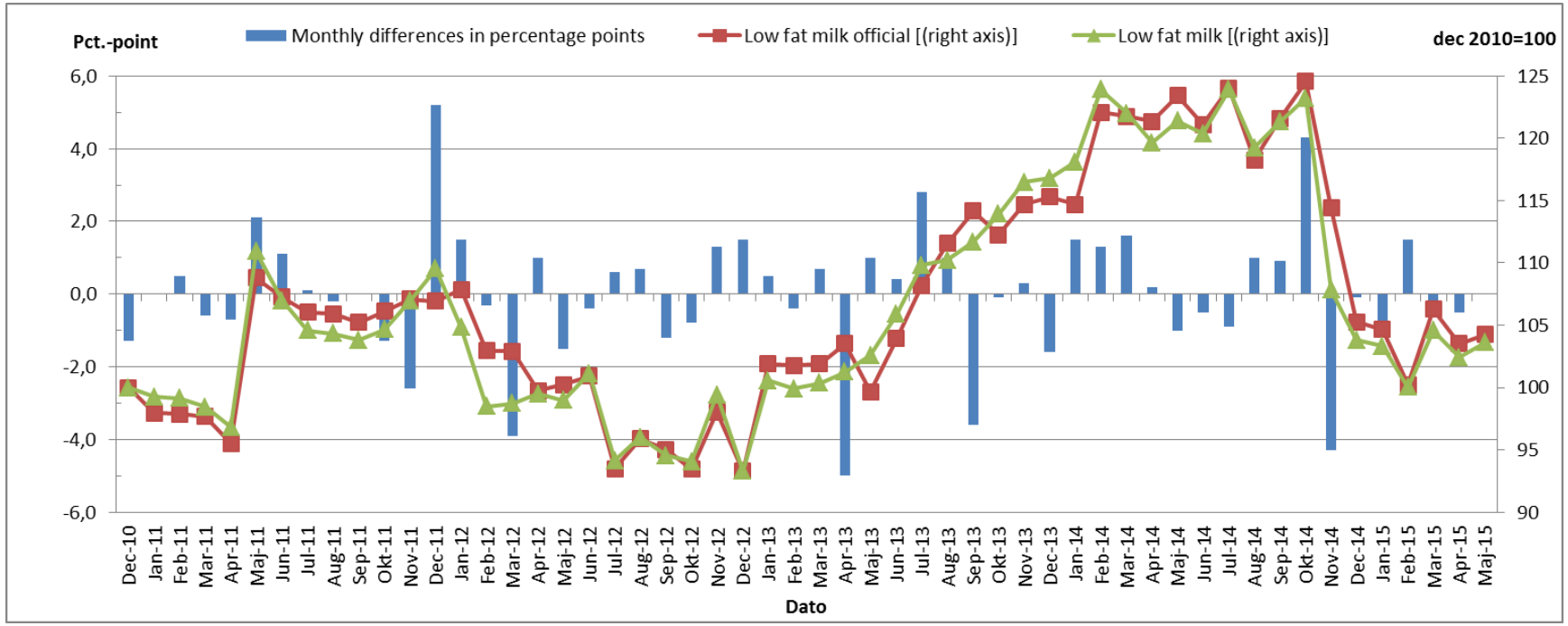
COICOP 1


Monthly differences in percentage points
Official [(right axis)]
Scanner data [(right axis)]
Pct.point
2000=100
2,0
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
-2,0
138
136
134
132
130
128
126
124
122
120
Dec-10
Jan-11
Feb-11
Mar-11
Apr-11
Maj-11
Jun-11
Jul-11
Aug-11
Sep-11
Okt-11
Nov-11
Dec-11
Jan-12
Feb-12
Mar-12
Apr-12
Maj-12
Jun-12
Jul-12
Aug-12
Sep-12
Okt-12
Nov-12
Dec-12
Jan-13
Feb-13
Mar-13
Apr-13
Maj-13
Jun-13
Jul-13
Aug-13
Sep-13
Okt-13
Nov-13
Dec-13
Jan-14
Feb-14
Mar-14
Apr-14
Maj-14
Jun-14
Jul-14
Aug-14
Sep-14
Okt-14
Nov-14
Dec-14
Jan-15
Feb-15
Mar-15
Apr-15
Maj-15

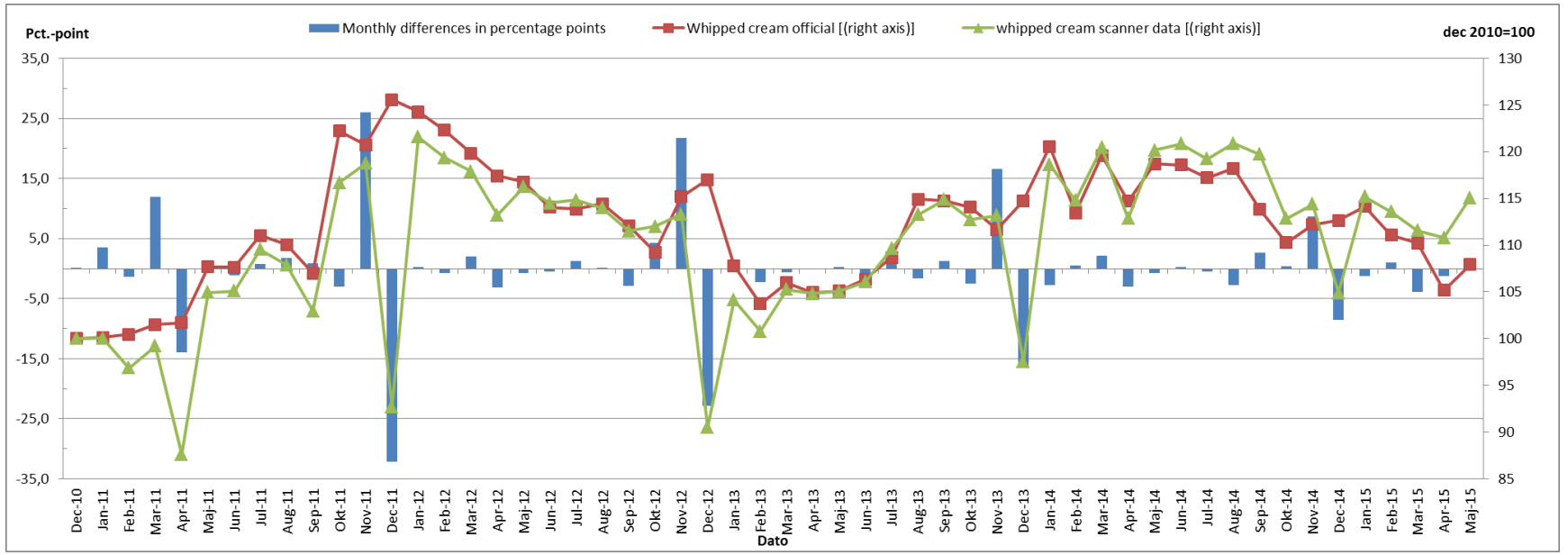
COICOP 2


Monthly differences in percentage points
Official [(right axis)]
Scanner data [(right axis)]
Pct.-point
2000=100
2,0
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
-2,0
140
135
130
125
120
115
110
Dec-10
Jan-11
Feb-11
Mar-11
Apr-11
Maj-11
Jun-11
Jul-11
Aug-11
Sep-11
Okt-11
Nov-11
Dec-11
Jan-12
Feb-12
Mar-12
Apr-12
Maj-12
Jun-12
Jul-12
Aug-12
Sep-12
Okt-12
Nov-12
Dec-12
Jan-13
Feb-13
Mar-13
Apr-13
Maj-13
Jun-13
Jul-13
Aug-13
Sep-13
Okt-13
Nov-13
Dec-13
Jan-14
Feb-14
Mar-14
Apr-14
Maj-14
Jun-14
Jul-14
Aug-14
Sep-14
Okt-14
Nov-14
Dec-14
Jan-15
Feb-15
Mar-15
Apr-15
Maj-15

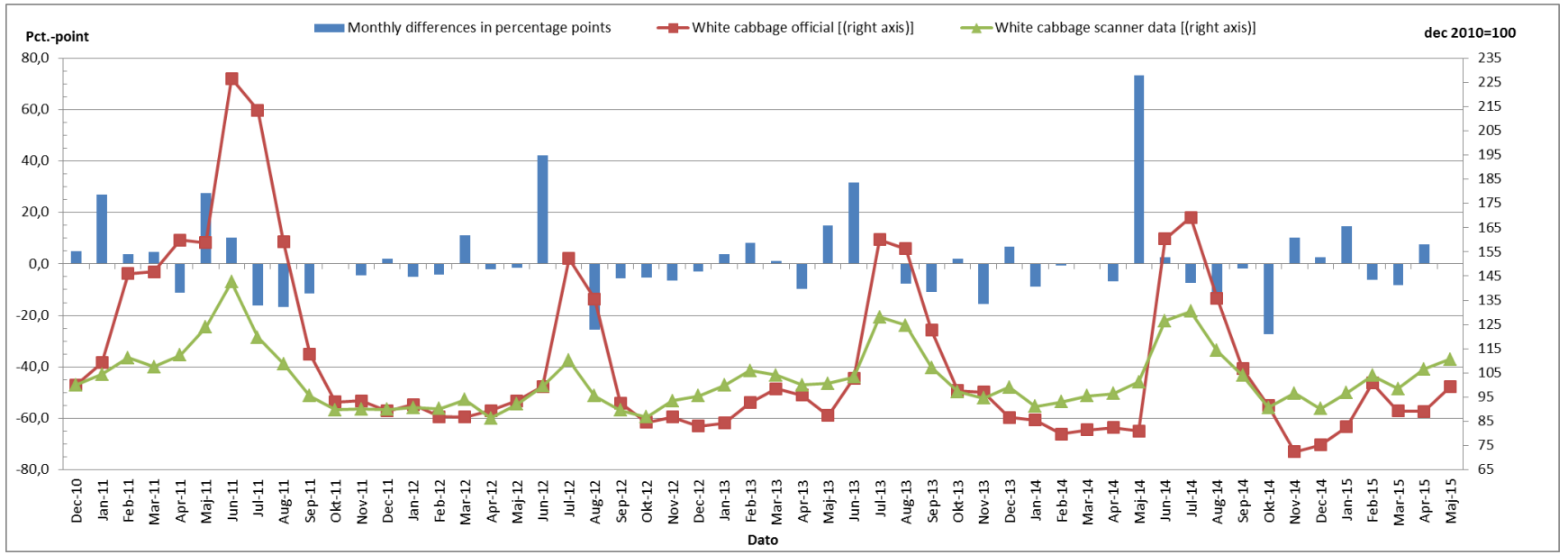
No significant differences



Scanner data is a better source



Scanner data is a different source



Worries

- The complexity of the data can involve the risk of a suboptimal sample
- Data is not delivered timely

Emergency systems

- We use two weeks for the calculation of SD-indices giving us a chance of switching weeks.
- Developed SAS-programs for the continuation of prices for one or more SD-chains
- We've signed contracts with the chains on the delivery of data. This will include a clause letting us know if the product structure changes

Lessons learnt

- 1. Scanner data is complex. Cleaning of data and an assumption of diversity of barcodes is necessary
- 2. The differences when handling scanner data as compared to traditional price collection need to be implemented in the IT-systems in a proper way and thought through as early as possible
- 3. Common frame work is important

On unit values

- In general unit values are applicable for the calculation of SD-indices

- But:
- The data needs to be cleaned in other variables before hand(product text, unit, amount…) at the lowest possible level
- The price development has to be monitored both within the month and between months

End discussion

- What are your worries with the processing of scanner data?