

Approaching the challenges with the new data eco-system

Karin Blix, Quality Coordinator, kwb@dst.dk Statistics Denmark







Challenges of the new data eco-system

Examples From Denmark

- Long tradition of using administrative data sources for official statistics.
- This presentation will elaborate on:
 - how the Danish statistical system relies on administrative sources and
 - how this is also a base for exploring new data sources and
 - what is done to ensure quality in official statistics in the new data eco-system.

Quality in statistics



 ISO, an international body for formulating standards, has defined quality as: "Degree to which a set of inherent characteristics fulfils requirements".

it for

- Quality in statistics is the set of properties the statistic needs to have to fulfil the users' needs.
- Our starting point is the users and their needs.

Administrative data





Data which are originally collected by public authorities for their own purposes

				COLORADO DE LA COLORADO				
	1 000 1	2 000 2 004 5 013	87 881 87 881 87 881 87 881 38 013 28 321 6 757 2 934 4 573	The second second				
Usually organised and structured in administrative registers								



Administrative registers are **a valuable asset** also for producing statistics

Data flow from data owners to end-users DENMARK

Data are collected in registers as by-products of administrative routines 5 & 6

Producing statistics

Creating knowledge Disseminating statistics to users in Society









Statistical information system





Examples of exploring data sources



Administrative data

- Life Lines is an example of exploring existing administrative sources for new uses
- Networks of opportunities another example exploring existing administrative sources

New data sources

- The use of scanner data for price statistics has been in regular use for many years
- AIS position data for ships is in use for experimental statistics on port calls
- Data from smart-meters on use of electricity is being explored for several uses.

Examples – use of admin data - population

Life-lines

- The longitudinal Register (Life lines) is updated annually and shows the individuals that have been part of the Danish population through their life courses.
- Based on the population statistics register: all persons in Denmark from 1986 to 2021 plus an extraction from the CPR (Central Person register) from 1968.
- One line for each person's presence in the population, e.g. as a period from birth to death or from the person's immigration to emigration/death
- The purpose: form a population based on life lines, e.g. extraction of cohort, period, age, length of life line, etc.



Examples - use of Life Lines for networks DENMARK

Network of opportunities

- Based on Life-Lines
- Networks composed of
- Nodes (individuals)
- Edges (relations between individuals)
- Degree (how many relations one individual has edges per node)
- Statistics on number of family members, colleagues, class mates etc.
- Segregation (how many one potentially knows that ha_ a university degree, two vacation homes, a mother from Norway etc.)
- How something can spread through a network (sickness, education, employment etc.)





Example – networks - descriptives

Descriptives





Example – networks

Family



Example – use of electricity data



- Database registry of construction projects
- Access to smart-meter electricity data on address level
- Machine learning estimation model correcting delay
- Quarterly statistics
- Issues
 - Late and incomplete registry regarding projects

New approach

 Combining electricity data with data about known construction projects

STATISTICS

 Exploring patterns of interest e.g. identifying construction phases

- Pilot project
 - Find a sample project and follow the path through the registry

Example – use of electricity data



Purchase

Smart-meter data: available with 8 days delay

New agreement from 15/06 appears in July Use of electricity begins on 18/7







Example - Future work - electricity data

Immediate potential

- Using electricity data as proxy
- Development of new indicator
- Now-casting statistic
- Pattern recognition for identifying more projects – data science lab
- Other uses of smart-meter data
 - Charging of electric cars
 - Identifying empty dwelling
 - Assessing use of summer homes
 - Forecasting bankruptcies



How we ensure quality

- Cooperation with data owners
- Cooperation with data providers (e.g. government agencies, municipalities, educational institutions etc.)
- Methods unit in Statistics Denmark
 - Sample selection
 - Standardised and modernised error detection
 - Other methodological issues
- Quality unit in Statistics Denmark
 - Quality assurance of documentation of statistics (quality reports)
 - Quality reviews
 - Standardisation e.g. the process model (GSBPM)
 - Coherent metadata system
- Data science lab in statistics Denmark
 - Exploring existing data sources
 - Exploring new data sources
 - Exploring new methodology





Statistical processes (GSBPM 5.1)



				Overarchi	ng Processes			
	Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
┍╸	1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
	1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
	1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
	1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
	1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
	1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
			3.7 Finalise production systems		5.7 Calculate aggregates			

5.8 Finalise data files

Production of statistics





Two types of metadata



Structural metadata

- Used to identify statistical data
- Headlines, variable names, unit of measure, reference time etc.
- <u>Must</u> go together with statistical data
- Impossible to interpret statistics without it

Reference metadata

- Describes content, statistical processing, relevance etc.
- <u>Can</u> be detached from the statistical output
- Quality Reports is a type of reference metadata
- ...so is methodological metadata

Statistics
without
metadata

2 881 620
2 908 337
2 868 172
2 976 785

with
structural
metadata

Population	
All Denmark	20180
Men	2 881 620
Women	2 908 337
Unit : number	

Real estate market value

One-family houses	2016
Brøndby	2 868 172
Vallensbæk	2 976 785
Unit · Average Market	value (DKK)

...and <u>reference</u> metadata

Population

All Denmark	:	2018Q3	
Men	2.8	81 620	
Women	2 9	08 337	
Unit : number	г	4	
Real estate marke	et va	DANMARKS STATISTIK	
One-family houses		Quality	Report
Brøndby	2		
Vallensbæk	2		
Unit : Average Market	value		

Documentation of statistics - content



STATISTICS

DENMARK

User profiles

General population

- Everyone in contact with SD through the flow of news
- Mr and Mrs Smith
- People interested in social affairs

Specially interested parties

- Actively searching for facts
- No special qualification
- E.g. journalists, students and politicians

Professional users

- Systematically use figures from SD
- Can combine and extract data
- E.g. specialists, trade and business press and public servants
- Analysts
 - Awareness of statistics
 - Can extract and process complex data
 - E.g. researchers, large-scale consumers and data analysis units







Frontpage > Find statistics > Documentation > Documentation of statistics > Climate footprint (experimental statistics)

^

← Documentation of statistics

Climate footprint (experimental statistics)
Statistical presentation
Statistical processing
Relevance
Accuracy and reliability
Timeliness and punctuality
Comparability

Accessibility and clarity

SHARE THIS PAGE X in \square

Climate footprint (experimental statistics)

The purpose of t	sta			Statbank	Contact Statistics Denmark	Information Service	Press	Dansk
gases from the :	STATISTICS							
investment and	DENMARK	Find statistics	Products & services		Consulting abroad	About us	S	earch O
emissions of gre		Find statistics	FTODUCTS & SETVICES		consutting abroad	About us		earch ~
Global emission	Frontpage > Find statistics > Documentation > Documentation of statist	tics > Climate footprint (experimental s	tatistics) > Statistical presentation	n				

Climate footprint (experimental

statistics)

Statistical pr Statistical presentation

The statistics sh Statistical processing supply chains fo

are distributed t Relevance by supplying ind

Accuracy and reliability Read more abou

Timeliness and punctuality Statistical pr

The climate foot Comparability Statistics Denm

Accessibility and clarity

SHARE THIS PAGE

M in

international pre Read more abou

constitutes Deni been prepared s which uses it for

Global Afrappor

Relevance

The climate foot global emission

Statistical presentation

The statistics show the amount of greenhouse gas that has been emitted in the supply chains for Danish final use annually from 1990 onwards. The emissions are distributed by type of final use, emitting industries and countries, as well as by supplying industries.

Data description

The statistics show the amount of greenhouse gas that has been emitted in the supply chains for Danish final use. The emissions are distributed by type of final use, emitting industries and countries, as well as by supplying industries.

The calculation of the climate footprint uses 100-year Global Warming Potentials from the IPCC's fourth assessment report (AR4) to convert tonnes of a given greenhouse gas into tonnes of CO2 equivalents.

The supply chain for a type of final use is defined in these statistics as all the production activities in Denmark and the rest of the world that have been necessary to produce the products for final use. The supply chain behind e.g. milk includes both raw milk production and further processing, the production of dairy cows and feed for them, the production of electricity to run the stables and dairies, as well as steel and wood to build the stables and dairies, etc.

The emissions are calculated in tonnes of CO2e (CO2 equivalents) and include the greenhouse gases CO2, CH4 (Methane), N2O (Nitrous oxide) and F-gases (SF6, HFC-gases and PFC-gases). In relation to LULUCF (unfccc int) the climate footprint only includes emissions from land use in the agricultural sector.

The statistic has five variables: - Types of use: The type of Danish final use whose supply chain led to the greenhouse gas emissions. - Supplying industry: The Danish industry that formed the last link in the supply chain for the final use. Imports from foreign industries directly for final use are entered under the item "Imports for final use" so that the statistics are fully comprehensive. (Supply industry is only included in the AFTRYK2 table) - Emitting industry. The industry where the production that emitted the greenhouse gas took place. - Emitting country: The country where the production that emitted the greenhouse gas took place. - Year: The year of final use.

Contact info

National Accounts, Economic Statistics Peter Rarmose Jensen «ul> «li>45 3917 3862 «/li> «/ul> prj@dst.dk

Get as PDF

Climate footprint (experimental statistics)

Vision: Metadata used in the entire production process



			Overarch	ng Processes			
Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluat
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evalue inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evalu
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an actior
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			



Input

- Surveys

- Validation
- Visual inspection
- Aggregation



Ristoriovaj ¥ fin ⊡	Lande i udenrigshandel med varer (GEONOM), v1:2021	
	Matth Laws - size equivable test care (18.0000) of 2011	
	Exploritation: Second and a CA 2000 per multiple to an address basis of enterior (permises of enteriors and Add Add Carl per multiple to any address of enterior (permises) 11 to a linearly explore and the	- WIT: 1 AMONE 1 ALT
	Syldig frec 1 production	- E1: EUROPA - F1: AFRIKA
	Autor management	
	Renalityenset lange activate to be charged as in the local	
		- F4. Nordafrika
	Form retransform	DZ: Algoriet
	Annual annual () Annual ()	EG: Egypten
	A Printerer	LY: Libven
	a to brack	Mar Marekke
	a for some	Per Parono
	· · · · · · · · · · · · · · · · · · ·	Inc Turresien
		 E2: Andre afrikanske lands

STATISTICS

DENMARK

Statbank tables



Conclusion

- At Statistics Denmark, we are constantly discovering new applications and gains from having a single system based approach to metadata reporting and storage.
- In combination with a dedicated and expanding team of quality practitioners principles of...
 - relevance,
 - accuracy & reliability,
 - timeliness & punctuality,
 - coherence & comparability and
 - accessibility & clarity
- ...are becoming routine and synonymous with the way we do business on a daily basis.



STATISTICS





Thank you for your attention

