



VOORBURG
GROUP

Voorburg Task Force - Alternative Data Sources

HELSINKI

24 | 25
SEPTEMBER
2020

Ana Aizcorbe

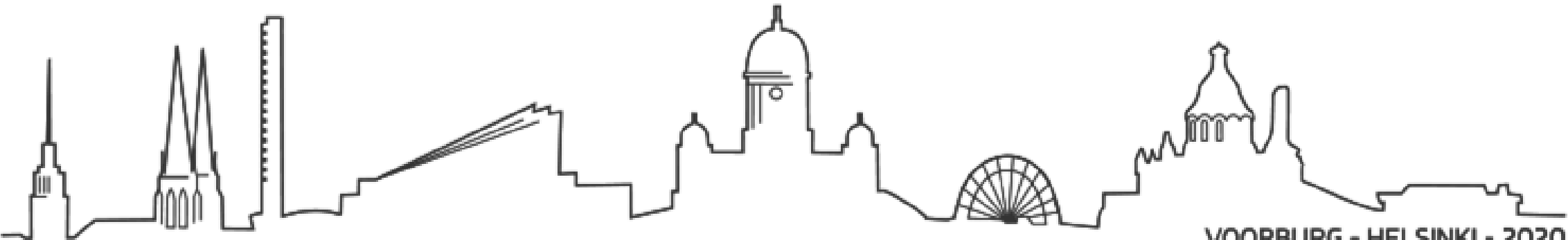
Rohan Draper

Nicklas Elversøe

Anna Korlyuk

Ville Markkanen

Susanna Tåg



VOORBURG - HELSINKI - 2020



Voorburg Task Force - Alternative Data Sources Foreword

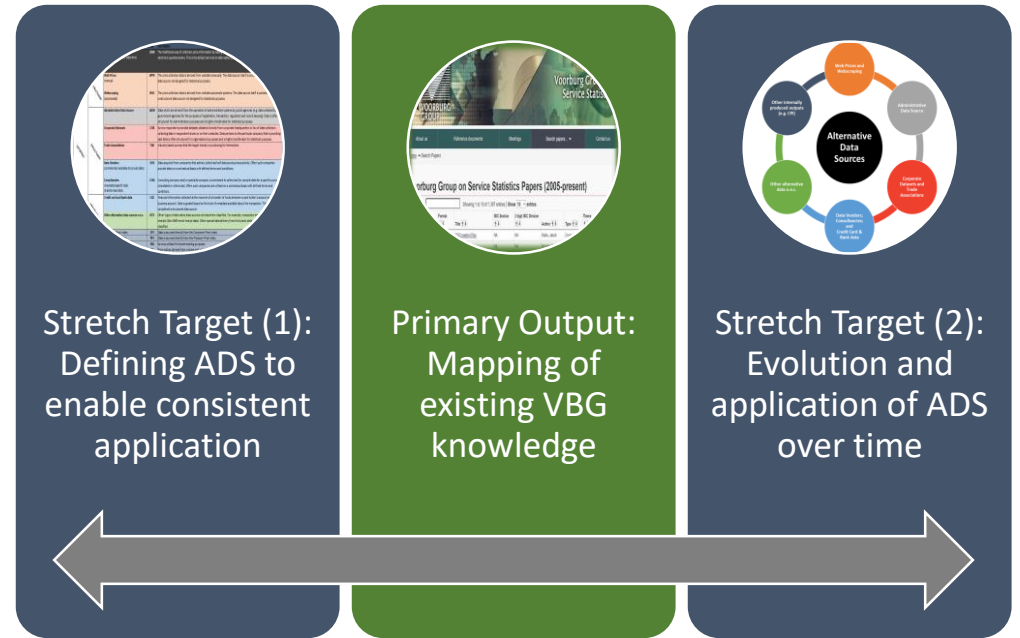
Foreword

This Voorburg Group Task Force was formed with the mandate of improving the Voorburg Group (VBG) awareness and accessibility of existing VBG documentation on alternative data sources.

The main repository of VBG knowledge is the VBG website and this has thus served as the primary source for the investigation. Secondary sources include the Ottawa Group website (the UN City Group responsible for international discussion and development of CPIs) and their respective documentation. The primary output from the Task Force is a mapping table of existing VBG documentation.

Early on in the process the Task Force discussed the apparent lack of foundation documentation and took on the stretch target of building a definition table covering alternative data sources (ADS) to support mapping of papers. Secondly, the Task Force discussed the sustainability of the work and its application in respective VBG members forward planning.

To this end, the Task Force took on a second stretch target of building a prototype questionnaire to enable the VBG to follow the growth in the breadth and depth of alternative data source use over time on a per 4-digit level basis. These two stretch targets are also presented in this report.





Voorburg Task Force - Alternative Data Sources Contents

Foreword	1
Definitions	3
Defining alternative data sources to enable consistent application	3
Method	5
Selection of papers from the VBG catalogue	5
Results	6
Tabulations by data source and year	6
Alternative data sources mostly for weighting purposes	6
Administrative data used for many industries	7
Questionnaire	8
Evolution and application of alternative data sources over time	8
Proposals and Discussion Items	9
Conclusion	10

Voorburg Task Force - Alternative Data Sources Definitions

HELSINKI
24 | 25
SEPTEMBER
2020

Defining alternative data sources to enable consistent application (stretch target 1)

In the pursuit of mapping current VBG documentation the Task Force wanted to enable price statisticians to quickly be able to determine if documentation was suitable for their respective needs. Beyond the industry and/or product classification a mapping to particular types of alternative data sources was determined to be of necessity. The challenge at this juncture was that no single document exists that ratifies definitions for alternative data sources and no clear foundation documentation exists for individual alternative data sources albeit perhaps web-scraping.

The definitions developed are largely based on a mix of documentation from the VBG website and Ottawa Group website. Ideally, harmonisation of principles and definitions across both PPI and CPI fronts would be ideal. Such definitions will at a minimum need to be ratified by the VBG and potentially also sent for discussion by the Ottawa Group.

Discussion (1): How does the Voorburg group feel about the proposed definitions? Are you aware of documentation that may be useful? Should we pursue contact with the Ottawa Group to harmonise definitions across both PPI and CPI fronts.

Source	Type	Code	Description
TRADITIONAL	STRUCTURED	Questionnaire (paper, phone and/or electronic)	QNR The traditional way of collection price information by asking firms for the information via telephone, paper or electronic questionnaires. This is the default and not an alternative data source.
	UNSTRUCTURED	Web Prices (manual)	WPR The price collection data is derived from websites manually. The data source itself is considered an unstructured data source not designed for statistical purposes.
Webscraping (automated)		WSC The price collection data is derived from websites automatic systems. The data source itself is considered an unstructured data source not designed for statistical purposes.	
EXTERNAL	STRUCTURED	Administrative Data Source	ADM Data which are derived from the operation of administrative systems by public agencies (e.g. data collected by government agencies for the purposes of registration, transaction, regulation and record keeping). Data is often structured for administrative purposes and is highly transferable for statistical purposes.
		Corporate Datasets	COR Survey respondent provided datasets obtained directly from corporate headquarters in lieu of data collectors collecting data in respondent stores or on their websites. Data pertains to the particular company that is providing said data is often structured for organisational purposes and is highly transferable for statistical purposes.
		Trade Associations	TAD Industry based surveys that the target industry is producing for themselves.
	UNSTRUCTURED	Data Vendors (commercially available structured data)	DVS Data acquired from companies that actively collect and sell data as a business activity. Often such companies provide data on a contractual basis with defined terms and conditions.
		Consultancies (mandated specific task) (transformed data)	CON Consulting company and/or specialist company is contracted to collect and/or compile data for a specific purpose (mandated or otherwise). Often such companies are utilised on a contractual basis with defined terms and conditions.
		Credit card and bank data	CCD Financial information collected at the moment of a transfer of funds between a card holder's account and a business account. Data is graded based on the level of metadata available about the transaction. This source is considered a structured data source.
BOTH	Other alternative data sources n.e.c.	OTH Other types of alternative data sources not elsewhere classified. For example, transaction-level data from email receipts (like UBER email receipt data). Other special data delivery from third party data collectors not elsewhere classified.	
INTERNAL	STRUCTURED	Consumer Price Index	CPI Data is sourced directly from the Consumer Price Index
		Producer Price Index	PPI Data is sourced directly from the Producer Price Index
		Structural Business Statistics	SBS Surveys utilised for benchmarking purposes
		National Accounts	NA Price indices derived from volume and value data (implicit price indices)

Note: An enlarged version of this table will follow on the next slide.

Source	Type	Code	Description	
EXTERNAL	TRADITIONAL	STRUCTURED	Questionnaire (paper, phone and/or electronic)	QNR The traditional way of collection price information by asking firms for the information via telephone, paper or electronic questionnaires. This is the default and not an alternative data source.
	UNSTRUCTURED	Web Prices (manual)	WPR	The price collection data is derived from websites manually. The data source itself is considered an unstructured data source not designed for statistical purposes.
		Webscraping (automated)	WSC	The price collection data is derived from websites automatic systems. The data source itself is considered an unstructured data source not designed for statistical purposes.
	STRUCTURED	Administrative Data Source	ADM	Data which are derived from the operation of administrative systems by public agencies (e.g. data collected by government agencies for the purposes of registration, transaction, regulation and record keeping). Data is often structured for administrative purposes and is highly transferable for statistical purposes.
		Corporate Datasets	COR	Survey respondent provided datasets obtained directly from corporate headquarters in lieu of data collectors collecting data in respondent stores or on their websites. Data pertains to the particular company that is providing said data is often structured for organisational purposes and is highly transferable for statistical purposes.
		Trade Associations	TAD	Industry based surveys that the target industry is producing for themselves.
		Data Vendors (commercially available structured data)	DVS	Data acquired from companies that actively collect and sell data as a business activity. Often such companies provide data on a contractual basis with defined terms and conditions.
		Consultancies (mandated specific task) (transformed data)	CON	Consulting company and/or specialist company is contracted to collect and/or compile data for a specific purpose (mandated or otherwise). Often such companies are utilised on a contractual basis with defined terms and conditions.
		Credit card and bank data	CCD	Financial information collected at the moment of a transfer of funds between a card holder's account and a business account. Data is graded based on the level of metadata available about the transaction. This source is considered a structured data source.
	BOTH	Other alternative data sources n.e.c.	OTH	Other types of alternative data sources not elsewhere classified. For example, transaction-level data from email receipts (like UBER email receipt data). Other special data delivery from third party data collectors not elsewhere classified.
INTERNAL	STRUCTURED	Consumer Price Index	CPI	Data is sourced directly from the Consumer Price Index
		Producer Price Index	PPI	Data is sourced directly from the Producer Price Index
		Structural Business Statistics	SBS	Surveys utilised for benchmarking purposes
		National Accounts	NA	Price indices derived from volume and value data (implicit price indices)

Voorburg Task Force - Alternative Data Sources Method (Mapping of existing VBG knowledge)

HELSINKI
24 | 25
SEPTEMBER
2020

Selection of Papers from the VBG catalogue

The Voorburg Group website provides links to documents collected from VBG conferences since 2005, including slide decks from presentations, posters and papers of different types. The following information is provided for each link: author, title, year and location of conference where the information was presented, format (paper, presentation, poster), Type (agenda, methodology, classification, country industry, cross cutting, issues paper, sector paper, task force) Theme (Output and Prices or Prices) and information on the topic. Often, each document is listed more than once: e.g., a paper with more than one author is listed twice, once under each author's name).

For papers presented from 2014 to 2019, we excluded papers listed as issues papers, cross cutting papers, or papers focused on classification issues; thinking the remaining papers would be most likely to provide detailed information about alternative data sources and how they are used. The remaining 42 papers were the focus of our analysis.



The screenshot shows the website header with the Voorburg Group logo and navigation links: About us, Reference documents, Meetings, Search papers..., and Contact us. Below the header is a search bar with the text "Home → Search Papers". The main content area is titled "Voorburg Group on Service Statistics Papers (2005-present)". It includes a filter items input field, a "Showing 1 to 10 of 1,397 entries" indicator, and a "Show 10 entries" dropdown. A table of search results is displayed with columns for Year, Location, Format, Title, ISIC Section, 2 digit ISIC Division, Author, Type, Theme, and Topic. The table shows two entries for the year 2019, both from Paris, with titles "2020 meeting Plan".

Year	Location	Format	Title	ISIC Section	2 digit ISIC Division	Author	Type	Theme	Topic
2019	Paris	Presentation	2020 meeting Plan	NA	NA	Kalko, Jakob	Cross-cutting	Other	General
2019	Paris	Presentation	2020 meeting Plan	NA	NA	Murphy, Bonnie	Cross-cutting	Other	General

Voorburg Task Force - Alternative Data Sources Results (Mapping of existing VBG knowledge)

HELSINKI
24 | 25
SEPTEMBER
2020

Tabulations by data source and year

Conference Year	2015	2016	2017	2018	2019	All years
Number of papers included in catalog	9	5	6	2	20	42
those using alternative data sources	4	1	3	1	7	16
Number of data sources considered						
Web Prices					2	2
Web-scraping						
Administrative Data Source	4	1	2	1	4	12
Corporate Datasets						
Trade Associations			1	1	1	3
Data Vendors			2	2	3	7
Consultancies					1	1
Other						
Total	4	1	5	4	11	25

Note: Because some papers analyzed more than one type of data source, totals for the rows in the bottom panel may exceed the number of papers that used the sources (e.g., 2018).

Of the 42 papers, 16 provided information on alternative data sources that were either being used in production or were being studied for potential use in the future. About half of those papers were presented in the latest VG conference (Paris), reflecting the increased interest in data sources in recent years. None of the papers used corporate datasets or automated web-scraping, though two papers discussed the use of websites as a source for price quotes. In total, there were 25 types data sources considered, the most common type being administrative data.

Alternative data sources mostly for weighting purposes

The construction of PPIs requires data with which to construct weights as well as price quotes and, often, data sources that can provide information on revenues that may be used for weights (typically annual or other low frequency data) do not have the high-frequency price quotes required for the index calculation. Most of the studies discussed potential data sources for weights (perhaps because that is what one typically finds in administrative data sources). Among the few that investigated alternate sources for price quotes, most considered were prices manually obtained from websites and data from data vendors.

Types of alternative data sources	Number of data sources	
	Weights	Prices
Web Prices		2
Administrative Data Source	11	1
Trade Associations	3	
Data Vendors	2	5
Consultancies	1	
Total	17	8

Voorburg Task Force - Alternative Data Sources Results (Mapping of existing VBG knowledge)

HELSINKI
24 | 25
SEPTEMBER
2020

Administrative data used for many industries

2 digit ISIC Division	Number of papers	Number of alternative data sources considered					Total
		Manual webscraping	Administrative Data Source	Trade Associations	Data Vendors	Consultancies	
H - Transportation and storage							
49 - Land transport	1		1		1		2
I - Accommodation							
55 - Accommodation Services	2	1	1	1			3
J - Information and communication							
61 - Telecomm Services	1		2				2
K - Financial and insurance activities							
64 - Financial service activities, except insurance and pension funding	4		2	2	5		9
65 - Insurance	1	1			1		2
66 - Activities auxiliary to financial service and insurance activities	1		1				1
73 - Advertising	1					1	1
M - Professional, scientific and technical activities							
74 - Other professional, scientific and technical activities	1		1				1
N - Administrative and support service activities							
79 - Travel agency, tour operator, reservation service and related activities	1		1				1
82 - Office administrative, office support and other business support activities	3		3				3
Total	16	2	12	3	7	1	25



Voorburg Task Force - Alternative Data Sources Questionnaire

Evolution and application of alternative data sources over time (stretch target 2)

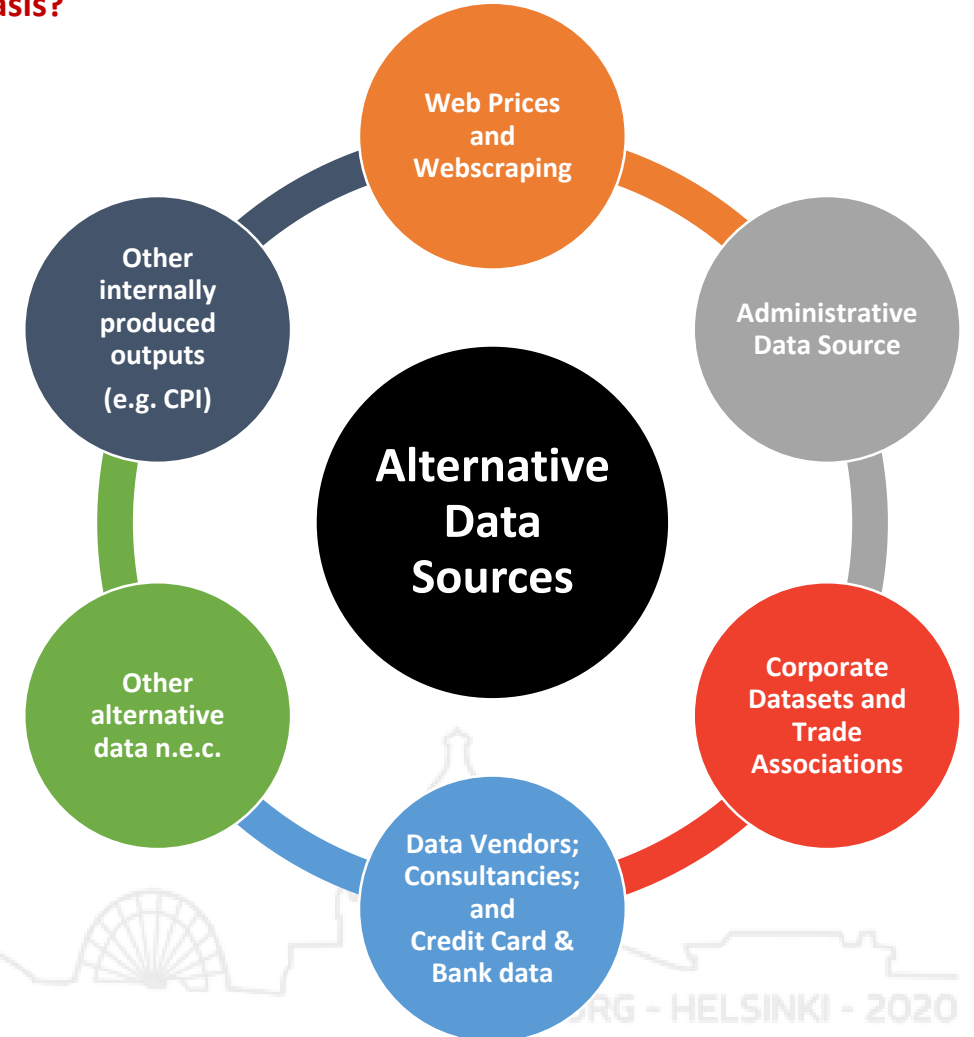
Concept

- Price statistics orientated mapping of data sources
- Map alternative data use per index for Voorburg members
- Enable Voorburg members to identify potential development partners
- Highlight trends in contemporary price index data source practises
- Improve decision making regarding SPPI development activities

How does the questionnaire function?

- Demo of the excel

Discussion (2): How does the Voorburg group feel about completing such an audit of data collection activities on an annual basis?





Voorburg Task Force - Alternative Data Sources Proposals and Discussion Items

Proposals

- **Timing and Frequency** – The Task Force recommends completion of the alternative data sources survey on an annual basis prior to the Voorburg session to be added to the country progress report.
- **Burden** – In order to reduce burden countries can upon request be sent their previous year's submission. It is perceived that after the initial year's completion updating the survey should require fairly minimal resources.
- **Availability** – The Task Force proposes that aggregate level data be made available via the Voorburg Group website but is open to suggestion from the broader group about exactly how this should function within regular Voorburg activities.
- **Future Papers** – When writing and presenting future papers the data source (or alternative data source) should be explicit and easy to locate. This may require an alteration to current VBG templates.

Discussion Items

- **Data Collection** – Should this survey be combined with the current country progress report or a stand alone data request?
- **Data Storage** – Currently the Voorburg Group Secretary maintains the master files for the country progress report. Does this seem like the best solution for the survey on alternative data sources too?
- **Ottawa Group (UN City Group: CPI)** – Would members of the Voorburg be comfortable with sharing aggregate level data with the Ottawa Group? Is this an aspect that we should be working together on?
- **SPPI Development Guide** – Should the guide be updated with information pertaining to alternative data sources?

Voorburg Task Force - Alternative Data Sources Conclusion

HELSINKI

24 | 25
SEPTEMBER
2020

Conclusion

Alternative data sources as a phenomenon is directly connected to the digitalization of economies. For some companies the goal is turning data into strategic insight for their own use and for others the focus is the monetization and commodification of data as a product for sale. Either way the change is rapid and our need to understand (and define) these alternative data sources of paramount importance and a distinct opportunity.

This Task Force has taken a first step in defining the types of alternative data that has been detected and documented by current VBG research and has put forward a proposal concerning its tracking for index development decision making.

Current documentation on alternative data sources is, however, ad hoc and very much in the developmental phase. There is an opportunity here to ratify current research; obtain consensus on definitions; and create a deliberate international agenda pertaining to the handling and application of alternative data sources. Example areas for examination include:

- A guideline on determining if an alternative data source is fit-for-purpose for use in the production of national statistics
- Legal ramifications when collecting data sets that aren't directly related to the business of the data provider
- How does industry levels of concentration and regulation impact opportunity to utilise alternative data sources
- How much are alternative data sources utilised for gap filling (coverage plugging) versus use as a primary source for index compilation

Two groups that naturally come to mind to spearhead such activities, at least on the price statistics front, are the Voorburg Group and the Ottawa Group but indeed this may need extension beyond just price statistics and as such other international groups may be worthwhile consulting as per their interest in being involved in such activities.