

Register-based Statistics

Statistical Methods for Administrative Data

- Anders Wallgren
- Britt Wallgren



Second Edition

Register-based Statistics

WILEY SERIES IN SURVEY METHODOLOGY

Established in Part by WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *Mick P. Couper, Graham Kalton, Lars Lyberg, J. N. K. Rao, Norbert Schwarz, Christopher Skinner*

A complete list of the titles in this series appears at the end of this volume.

Register-based Statistics

Statistical Methods for Administrative Data

Second Edition

Anders Wallgren and Britt Wallgren
*Formerly of the Department of Research and
Development at Statistics Sweden*

WILEY

This edition first published 2014
© 2014 John Wiley & Sons, Ltd

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Wallgren, Anders, author.

Register-based statistics : statistical methods for administrative data / Anders Wallgren and Britt Wallgren. – Second edition.

pages cm.

Includes bibliographical references and index.

ISBN 978-1-119-94213-9 (cloth)

1. Register-based statistics. I. Wallgren, Britt, author. II. Title.

HA31.23.W35 2014

519.5–dc23

2014003205

A catalogue record for this book is available from the British Library.

ISBN: 978-1-119-94213-9

Set in Times New Roman 11/12 pt by the authors.

Contents

Preface		xi
Chapter 1	Register Surveys – An Introduction	1
1.1	The purpose of the book	1
1.2	The need for a new theory and new methods	3
1.3	Four ways of using administrative registers	5
1.4	Preconditions for register-based statistics	6
	1.4.1 <i>Reliable administrative systems</i>	7
	1.4.2 <i>Legal base and public approval</i>	8
1.5	Basic concepts and terms	10
	1.5.1 <i>What is a statistical survey?</i>	10
	1.5.2 <i>What is a register?</i>	11
	1.5.3 <i>What is a register survey?</i>	13
	1.5.4 <i>The Income and Taxation Register</i>	14
	1.5.5 <i>The Quarterly and Annual Pay Registers</i>	16
1.6	Comparing sample surveys and register surveys	20
1.7	Conclusions	23
Chapter 2	The Nature of Administrative Data	25
2.1	Different kinds of administrative data	25
2.2	How are data recorded?	26
2.3	Administrative and statistical information systems	27
2.4	Measurement errors in statistical and administrative data	29
2.5	Why use administrative data for statistics?	30
2.6	Comparing sample survey and administrative data	32
	2.6.1 <i>A questionnaire to persons compared with register data</i>	32
	2.6.2 <i>An enterprise questionnaire compared with register data</i>	34
2.7	Conclusions	36
Chapter 3	Protection of Privacy and Confidentiality	37
3.1	Internal security	38
	3.1.1 <i>No text in output databases</i>	38
	3.1.2 <i>Existence of identity numbers</i>	39
3.2	Disclosure risks – tables	41
	3.2.1 <i>Rules for tables with counts, totals and mean values</i>	41
	3.2.2 <i>The threshold rule – analyse complete tables</i>	43
	3.2.3 <i>Frequency tables are often misunderstood</i>	44

	3.2.4	<i>Combining tables can cause disclosure</i>	45
	3.3	Disclosure risks – microdata	45
	3.4	Conclusions	46
Chapter 4		The Register System	47
	4.1	A register model based on object types and relations	47
		4.1.1 <i>The register system and protection of privacy</i>	53
		4.1.2 <i>The register system and data warehousing</i>	53
	4.2	Organising the work with the system	54
	4.3	The populations in the system	56
		4.3.1 <i>How to produce consistent register-based statistics</i>	57
		4.3.2 <i>Registers and time</i>	58
		4.3.3 <i>Populations, variables and time</i>	59
	4.4	The variables in the system	60
		4.4.1 <i>Standardised variables in the register system</i>	60
		4.4.2 <i>Derived variables</i>	62
		4.4.3 <i>Variables with different origins</i>	63
		4.4.4 <i>Variables with different functions in the system</i>	64
	4.5	Using the system for micro integration	65
	4.6	Three kinds of registers with different roles	70
	4.7	Register systems and register surveys within enterprises	72
	4.8	Conclusions	74
Chapter 5		The Base Registers in the System	77
	5.1	Characteristics of a base register	77
	5.2	Requirements for base registers	78
		5.2.1 <i>Defining and deriving statistical units</i>	78
		5.2.2 <i>Objects and identities – requirements for a base register</i>	80
		5.2.3 <i>Coverage and spanning variables in base registers</i>	81
	5.3	The Population Register	83
	5.4	The Business Register	88
	5.5	The Real Estate Register	93
	5.6	The Activity Register	94
	5.7	Everyone should support the base registers	98
	5.8	Conclusions	101
Chapter 6		How to Create a Register – Matching and Combining Sources	103
	6.1	Preconditions in different countries	103
	6.2	Matching methods and problems	105
		6.2.1 <i>Deterministic record linkage</i>	105
		6.2.2 <i>Probabilistic record linkage</i>	106
		6.2.3 <i>Four causes of matching errors</i>	112
	6.3	Matching sources with different object types	114
	6.4	Conclusions	120
Chapter 7		How to Create a Register – The Population	121
	7.1	How should register surveys be structured?	121
	7.2	Register survey design	125
		7.2.1 <i>Determining the research objectives</i>	125

	7.2.2	<i>Making an inventory of different sources</i>	128
	7.2.3	<i>Analysing the usability of administrative sources</i>	128
7.3		Defining a register's object set	131
	7.3.1	<i>Defining a population</i>	131
	7.3.2	<i>Can you alter data from the National Tax Agency?</i>	134
	7.3.3	<i>Defining a population – primary registers</i>	135
	7.3.4	<i>Defining a population – integrated registers</i>	136
	7.3.5	<i>Defining a calendar year population</i>	137
	7.3.6	<i>Defining a population – frame or register population?</i>	138
	7.3.7	<i>Base registers should be used when defining populations</i>	141
7.4		Defining the statistical units	142
	7.4.1	<i>Units and identities when creating primary registers</i>	143
	7.4.2	<i>Using administrative objects instead of statistical units</i>	144
7.5		Creating longitudinal registers – the population	145
7.6		Conclusions	146
Chapter 8		How to Create a Register – The Variables	147
8.1		The variables in the register	147
	8.1.1	<i>Variable definitions</i>	148
	8.1.2	<i>Variables in statistical science</i>	149
	8.1.3	<i>Variables in informatics</i>	150
	8.1.4	<i>Creating register variables – checklist</i>	151
8.2		Forming derived variables using models	151
	8.2.1	<i>Exact calculation of values using a rule</i>	152
	8.2.2	<i>Estimating values with a rule</i>	153
	8.2.3	<i>Estimating values with a causal model</i>	154
	8.2.4	<i>Derived variables and imputed variable values</i>	157
	8.2.5	<i>Creating variables by coding</i>	158
8.3		Activity data	159
	8.3.1	<i>Activity statistics</i>	160
	8.3.2	<i>Activity data aggregated for enterprises and organisations</i>	161
	8.3.3	<i>Activity data aggregated for persons: multi-valued variables</i>	161
8.4		Creating longitudinal registers – the variables	165
8.5		Conclusions	169
Chapter 9		How to Create a Register – Editing	171
9.1		Editing register data	171
	9.1.1	<i>Editing one administrative register</i>	173
	9.1.2	<i>Consistency editing – is the population correct?</i>	175
	9.1.3	<i>Consistency editing – are the units correct?</i>	178
	9.1.4	<i>Consistency editing – are the variables correct?</i>	180
9.2		Case studies – editing register data	181
	9.2.1	<i>Editing work within the Income and Taxation Register</i>	181
	9.2.2	<i>Editing work with the Income Statement Register</i>	183
	9.2.3	<i>What more can be learned from these examples?</i>	184
9.3		Editing, quality assurance and survey design	185
	9.3.1	<i>Survey design in a register-based production system</i>	185
	9.3.2	<i>Quality assessment in a register-based production system</i>	186
	9.3.3	<i>Total survey error in a register-based production system</i>	191
9.4		Conclusions	192

Chapter 10	Metadata	193
10.1	Primary registers – the need for metadata	193
	10.1.1 <i>Documentation of administrative sources</i>	194
	10.1.2 <i>Documentation of sources within the system</i>	194
	10.1.3 <i>Documentation of a new register</i>	195
10.2	Changes over time – the need for metadata	195
10.3	Integrated registers – the need for metadata	196
10.4	Classification and definitions database	197
10.5	The need for metadata for registers	198
10.6	Conclusions	200
Chapter 11	Estimation Methods – Introduction	201
11.1	Estimation in sample surveys and register surveys	202
11.2	Estimation methods for register surveys that use weights	203
11.3	Calibration of weights in register surveys	204
11.4	Using weights for estimation	207
11.5	Conclusions	208
Chapter 12	Estimation Methods – Missing Values	209
12.1	Make no adjustments, publish ‘value unknown’	210
12.2	Adjustment for missing values using weights	214
12.3	Adjustment for missing values by imputation	215
12.4	Missing values in a system of registers	218
12.5	Conclusions	220
Chapter 13	Estimation Methods – Coverage Problems	221
13.1	Reducing overcoverage and undercoverage	221
	13.1.1 <i>Coverage problems in the Population Register</i>	221
	13.1.2 <i>Coverage problems in the Business Register</i>	222
13.2	Estimation methods to correct for overcoverage	224
13.3	Undercoverage in the administrative system	226
13.4	Conclusions	228
Chapter 14	Estimation Methods – Multi-valued Variables	229
14.1	Multi-valued variables	229
14.2	Estimation methods	232
	14.2.1 <i>Occupation in the Activity and Occupation Registers</i>	232
	14.2.2 <i>Industrial classification in the Business Register</i>	236
	14.2.3 <i>Importing many multi-valued variables</i>	238
	14.2.4 <i>Consistency between estimates from different registers</i>	242
	14.2.5 <i>Multi-valued variables – what is done in practice?</i>	245
	14.2.6 <i>Additional estimation methods</i>	247
14.3	Application of the method	251
14.4	Linking of time series using combination objects	254
	14.4.1 <i>Linking time series</i>	254
	14.4.2 <i>Changed industrial classification in the Business Register</i>	256
14.5	Conclusions	258

Chapter 15	Theory and Quality of Register-based Statistics	259
15.1	Is there a theory for register surveys?	259
	15.1.1 <i>Statistical inference at a national statistical office</i>	260
	15.1.2 <i>Theory-based methods or ad hoc methods</i>	262
	15.1.3 <i>The survey approach and the systems approach</i>	263
15.2	Measuring quality – why and how?	267
15.3	Analysing administrative sources – input data quality	271
15.4	Output data quality	278
15.5	The integration process – integration errors	279
	15.5.1 <i>Creating register populations – coverage errors</i>	280
	15.5.2 <i>Creating statistical units – errors in units</i>	282
	15.5.3 <i>Creating statistical variables – errors in variables</i>	283
15.6	Random variation in register data	288
15.7	The register system and data warehousing	291
15.8	Conclusions	295
Chapter 16	Conclusions	297
	References	301
	Index	307

Preface

From the preface to the first edition

Register surveys are becoming increasingly common within a growing number of national statistical offices. However, they are also common within enterprises and other organisations, where data from the organisation's own administrative systems are used to produce statistics on, for example, production, sales and wages.

Although register-based statistics are the most common form of statistics, no well-established theory in the field has existed up to now. There have been no well-known terms or principles, which have made the development of both register-based statistics and register-statistical methodology all the more difficult. As a consequence of this, ad hoc methods have been used instead of methods based on a generally accepted theory.

Many countries are investigating the possibilities to use an increasing amount of administrative data for statistical purposes. It is necessary to reduce response burden and costs; increasing nonresponse in censuses and sample surveys also makes this new strategy necessary. A new approach is necessary and register surveys require that suitable statistical methods be developed.

We have studied the requirements for register-based statistics through analysis of Statistics Sweden's system of statistical registers. Since 1994, we have devoted an increasing part of our work, at the Department of Research and Development at Statistics Sweden, to the study of register surveys. We have also worked together with a number of manufacturing enterprises and analysed their administrative data for the purposes of management. These experiences are also used in this book.

The first version of this book was published in 2004 in Swedish. It has been used in a number of study groups within Statistics Sweden. Around 50 people at Statistics Sweden have read and commented on different parts of the first Swedish version of this book. In addition, several individuals were interviewed to provide material for different examples and methodological sections.

The study groups based on the Swedish book gave us a very good overview of methodological problems regarding the register-based statistics produced by Statistics Sweden and helped us in our work with the first edition of the English version that was published in 2007.

Our work on the second edition

We have used the first edition in a number of courses given in Europe and Latin America. The first edition was translated into Spanish by INEGI, the national statistical office in Mexico. It was very important for us to have the opportunity to discuss register-based statistics with colleagues from Latin America and learn

about their quite different preconditions regarding administrative data and statistics production. Our experiences from these courses and discussions have been incorporated in the new edition.

Since 2010 we have worked together with Professor Thomas Laitila at Örebro University. He has inspired us to think about the entire production system at a national statistical office. In the first edition we mainly discussed the register system, but in the second edition we also discuss the production system as a whole. Together with Thomas Laitila, we have worked with a research project regarding the quality of administrative data for economic statistics. The main results of this project are used in the new edition.

Our supporters and sources of inspiration

Our work with register-based statistics at Statistics Sweden was supported by Jan Carling, Director General 1993–1999, and Svante Öberg, Director General 1999–2005. Their active support was necessary for the success of our work.

Our courses in Latin America have been sponsored by the Inter-American Development Bank (IDB) and the United Nations Population Fund (UNFPA). The Spanish translation of the first edition was sponsored by the IDB. Finally, the research project on the quality of administrative data for economic statistics was a part of the BLUE-ETS project financed by the European Commission. Thanks to these sponsors, we have acquired experiences that have been very important for our work on the second edition.

Professor Carl-Erik Särndal has been a very important discussion partner during our work on the book. We have discussed important and difficult issues with him from the beginning of our work with the Swedish version to when we completed the second English edition. His broad experience from statistical offices in different countries and his background as a specialist in sample surveys have been enormously useful.

It is our hope that *Register-based Statistics – Statistical Methods for Administrative Data* and its proposals will stimulate the discussion of register statistics and give support to those who work with administrative data at national statistical offices.

Örebro, Sweden

Anders Wallgren
Britt Wallgren
ba.statistik@telia.com

Register Surveys – An Introduction

Three types of statistics based on microdata are published by national statistical offices – statistics based on *sample surveys*, statistics based on *censuses* and statistics based on *administrative registers*. This book deals with the third type, statistics based on administrative registers, where instead of collecting data through sample surveys and censuses, administrative registers from different sources are adapted and processed to make the data suitable for statistical purposes. This kind of survey is called a *register survey*.

We introduce a number of concepts and principles that are used when discussing register surveys. These concepts and principles form the basis for a theory of this type of survey. We primarily discuss register surveys at national statistical offices. There is growing interest in this area; many countries increasingly use administrative data for statistical purposes, and there is a growing demand for a theory of register surveys.

1.1 The purpose of the book

Our main purpose is to describe and explain the methods that should be used for register surveys. Conducting a register survey means that a new *statistical register* is created with existing sources. The statistical register is then used to produce estimates required for the survey. What methods should be used in creating such a statistical register? One or more administrative registers are used when a new statistical register is created and the statistical register can differ from the administrative sources in many ways.

A *system of statistical registers* consists of a number of registers that can be linked to each other. In the Nordic countries, the national statistical offices have developed systems of registers that are used in the production of statistics. When new statistical registers are created, this register system becomes an important source that can be used together with different administrative sources. Another purpose of the book is to explain how such register systems should be designed and used in the production of statistics.

When a national statistical office starts using more and more administrative sources, the *statistical production system* of that office will gradually change. From a system based on enumerators or interviewers, address lists or maps, the system will become increasingly register-based. Sample surveys will be based on the

Population Register or the Business Register instead of address lists or maps – variables in sample surveys can come from administrative registers as well as from telephone interviews or questionnaires. In addition to the change in methods used for sample surveys, new kinds of register-based statistics can also be produced. A third purpose of the book is to explain how administrative registers can be used to change the statistical production system of a national statistical office to improve cost efficiency and statistical quality.

Preconditions in different countries

The Nordic countries started to use administrative registers during the 1960s when paper-based administrative registers were transformed into computer-based flat files. The preconditions for using administrative registers for statistical purposes were good. This explains why the Nordic statistical offices now have access to large amounts of administrative data,¹ and that the quality of these data is high in comparison with most other countries. Consequently, it has been possible to create statistical register systems that have made statistics production efficient and even to conduct completely register-based population and housing censuses. Identifying variables as identity numbers for persons and enterprises have high quality and deterministic matching is therefore easy.

The preconditions for using administrative data in many countries are today not as good, and changing the production system into a register-based system will take many years. During that period, administrative systems will gradually be improved, so many other countries will be able to use administrative data efficiently in the future. Therefore, a clear understanding of the Nordic experiences from the beginning will facilitate development in new register countries.

However, we also discuss problems that arise in statistical offices in countries without the same preconditions. In North America, there is another tradition of working with administrative data. When identifying variables are of lower quality and coverage of administrative systems is poorer, methods have been developed for linking records and estimating population size that are important to use under these circumstances.

Our aim is to present statistical methods and principles of general interest, and we rely mostly on experiences and case studies from Statistics Sweden to illustrate these general methodological issues. As a complement to this aim, we also present some cases from new register countries that have recently started to develop register-based statistics.

We started writing books on register-based statistics during the 1990s, and during these years we have had access to registers and colleagues at Statistics Sweden. This access to a fully register-based production system has been vital for analysing and discussing register-based statistics.

Case studies are essential – in a book on register-based statistics we cannot present ideas with formulas as in books on sampling theory. We use case studies based on real data and charts with small miniature registers to illustrate register-statistical methods and quality issues.

¹ About 99% of the microdata stored in Statistics Sweden's databases come from administrative registers.

1.2 The need for a new theory and new methods

Sample surveys are based on methods that have been derived from an established theory – *sampling theory*. This theory has been developed within the academic world and statistical offices, and consists of terms and principles that are generally well known. Scientific literature and journals develop and spread the methodologies for sampling and estimation. Because the terms and principles are well known, people working with sample surveys can easily communicate and exchange their experiences.

Censuses with their own data collection are based on a long tradition of population censuses and the collection of data from local authorities, schools and enterprises. Measurement errors, design of questionnaires and nonresponse are methodological issues that also apply to sample surveys. Censuses and sample surveys are closely related in terms of methodology – censuses are often considered as special cases where the sample is the entire population.

Although register-based statistics are a common form of statistics used for official statistics and business reports, no well-established theory in the field exists. There are no recognised terms or principles, which makes the development of register-based statistics and register-statistical methodology all the more difficult. As a consequence, *ad hoc methods are used instead of methods based on a generally accepted theory*.

One important reason for this shortfall is that the subject field of register surveys is not included in academic statistics. Statistical theory within statistical science is understood as consisting of probability theory and statistical inference. Sampling theory is included within this theoretical school of thought, but register surveys based on total enumeration are not.

Unfortunately, statistical science has so far not included any theory on statistical systems. Statistical offices, larger enterprises and organisations do not often carry out separate surveys. It is more common that *statistical information systems* are built, which constantly generate new data. A statistical theory is necessary to describe the general principles and to develop the conceptual apparatus for such statistical information systems. Register surveys should be included in this theory. We formulate four basic principles for using administrative registers (Chart 1.1).

Chart 1.1 Four principles for using administrative registers for statistics

Transformation principle

Administrative registers should be transformed into statistical registers.
All relevant sources should be used and combined during this transformation.

System principle

All statistical registers should be included in a coordinated register system.
This system will ensure that all data can be integrated and used effectively.

Consistency principle

Consistency regarding populations and variables is necessary for the coherence of estimates from different register surveys.

Quality principle

The register system should be used for quality assessment of statistical surveys based on microdata comparisons with other surveys in the production system.

We use these principles in the book and gradually introduce the register-statistical terms that are needed for the discussions.

Chart 1.2 illustrates the present situation. Estimates from four different surveys are compared, and these comparisons show clearly that the systems approach often is missing in the work with statistical surveys. People are fully occupied with their own surveys and different surveys are also published at different points in time. As a rule most estimates are unique for one survey, but in Chart 1.2 we have found one identical variable and created the table with corresponding estimates from each survey. If we look at one survey at a time, we do not see any errors except for the sample survey in (4) where we have margins for the sampling error. But when we look at the four surveys together, we understand that there must be more serious errors in these surveys. We thus need *a theory for systems of surveys and new methods for quality assessment*. We return to this example in later chapters.

Chart 1.2 Employees by economic activity, November 2004, thousands

Economic activity	Business Register		Employment Register	Labour Force Survey	Survey Error margin
	Enterprises	Local units			
	(1)	(2)	(3)	(4)	(5)
Agriculture, forestry, fishing	35	37	37	26	5
Mining, quarrying, manufacturing	688	636	717	640	23
Electricity, gas and water	21	22	28	29	5
Construction	197	209	215	199	14
Wholesale and retail trade	456	453	484	456	20
Hotels and restaurants	89	93	99	106	10
Transport, communication	240	242	243	236	15
Financial intermediation	83	77	85	78	9
Real estate, business activities	457	524	457	470	20
Government	139	215	239	230	15
Education	382	408	431	462	20
Health and social work	836	684	675	675	24
Other service activities	142	163	175	168	13
Unknown activity	0	0	38	4	
Total	3 763	3 763	3 924	3 778	43

Why are there such large differences between the surveys? The estimates for mining, quarrying and manufacturing can be 636 or 717 thousands – the inconsistencies are more serious than the sampling error. The methodological work should consist of three steps: compare surveys and find errors and inconsistencies; find out why we have these inconsistencies; and finally, reduce the errors and inconsistencies.

Chart 1.2 also illustrates that we only have one established way of giving a numerical description of the quality of published estimates – margins for the sampling error. There is no commonly used way of describing the quality of register-based statistics. However, the non-sampling errors of sample surveys are as a rule not described in the same clear manner as the sampling errors; here we also lack methods for giving a numerical description of the quality of published estimates.

In 1995, Statistics Denmark published *Statistics on Persons in Denmark – A Register-based Statistical System*. The Danish book presents a systematic review of register-statistical work and describes how to design a well-prepared register system. The book was the first attempt to create a theory for register-based statistics and to describe the methods that are used. We build on and add to that work in this book.

1.3 Four ways of using administrative registers

When a statistical office plans to use administrative registers for statistical purposes, the office faces a survey design issue. How should the new sources be used? How should the existing surveys be modified or reduced? To answer these questions the administrative sources should be analysed by experienced subject-matter specialists and methodologists with a good overview of the production system.

An administrative register or source can be used in four different ways:

1. *Completely alone.*

If the source has good coverage and the variables in the source are of good quality, then the source can be used alone for producing statistics. Trade statistics based on only administrative registers with monthly data from Customs are an example of a source that many countries use alone for statistics production.

2. *Alone, but combined with a base register.*

The Population Register and the Business Register are two important base registers that are used for all surveys regarding persons or enterprises in the Nordic countries. Base registers are discussed in Chapter 5. If an administrative register or source is combined with a base register, the quality can be improved and controlled. It will then be possible to produce consistent register-based statistics.

The base register contains important classification variables that can be used together with the administrative source. The Annual Pay Register in Section 1.5.4 is an example of using a source in this way.

3. *In combination with a base register and other administrative registers.*

In many cases an administrative register does not have sufficient coverage and the variable content is too limited. Then it is not advisable to use the source alone for statistics production. But if many sources are combined, it may often be possible to use the combined data set to produce register-based statistics. We mention two examples of this kind.

Example: In the Swedish Income and Taxation Register of persons, about 30 different sources are used regarding different kinds of income. If all these different kinds of income are combined, it is possible to create *disposable income* of good quality for all persons.

Example: A business register at a national statistical office is based on administrative sources. With five sources we created a Business Register for Sweden containing all enterprises active during a specific year. Each source consists of the legal units in one taxation system. In the table below, undercoverage and overcoverage of the sources are compared with our final Business Register. The

administrative object sets in each source are adequate for each of the five taxation systems. Taken alone, each source is of low statistical quality; however, if all sources are combined, the coverage is good.

Over- and undercoverage in five administrative sources, per cent of all legal units

	Source 1	Source 2	Source 3	Source 4	Source 5
Overcoverage	41%	0%	0%	0%	0%
Undercoverage	21%	74%	74%	30%	9%

4. *To improve other surveys, i.e. to improve the production system.*

Example: There was no information on economic activity for some small enterprises in the Business Register. In the yearly income tax returns from small enterprises, there is text information from the enterprise that describes economic activity. This text was automatically coded into economic activity. In this way the yearly income tax returns were used to improve the Business Register.

In the Nordic countries, most register surveys use a base register as in 2 and 3 above. New register countries that have not yet developed good base registers will start with register surveys of the simple kind as in 1 above. When base registers have been developed, it will be possible to create register surveys according to 2 and 3.

1.4 Preconditions for register-based statistics

Preconditions differ between countries for sample surveys, censuses and register surveys; hence, the preconditions for statistical methods are different. The choice between cluster sampling and one-stage sampling depends on whether you have a Population Register or if you must use address lists. Regression estimation and calibration are methods that depend on the number and quality of available register variables. This means that an *increased use of administrative registers will change the preconditions for all kinds of surveys.*

For register surveys, the differences between countries are even more significant. Legislation on national registration and the taxation of persons and enterprises determine the character of the administrative systems that are used in each country. The legislation regarding statistical production and protection of statistical data also differs, and as a consequence certain methodological issues are important in some countries but not in others. The two main preconditions for using administrative registers for statistical purposes are stated in Chart 1.3.

Chart 1.3 Two preconditions for using administrative registers for statistics

Identity number principle

Unified systems of identity numbers are used in all administrative systems. The same identity number should follow an object over its lifetime.

Legal principle

A statistical office should have access to administrative registers kept by public authorities. This right should be supported by law and the protection of privacy must also be protected by law.

1.4.1 Reliable administrative systems

Reliable administrative systems will generate data of good administrative quality. Good administrative quality is a necessary but not sufficient condition for good statistical quality. The systems for tax administration and welfare programmes will gradually develop and change, and these changes will determine what administrative data can be used for statistical purposes in the future. It is therefore important that national statistical offices maintain close and long-term relations with administrative authorities and politicians.

The long-term strategy requires high-level contacts to promote strategic changes that will improve statistics production. The statistical office must explain to the administrative authorities how their data are used for statistical purposes. The statistical office also needs detailed information on how the administrative systems are organised and what changes are planned. Close and long-term contacts at all levels are required for these purposes.

What aspects of national administrative systems are important for statistical offices? We note two such aspects here, coverage and identity codes.

Coverage – the systems should cover all

The Nordic systems for child benefits are good examples. All children in defined age groups are entitled to a sum of money. All parents want the entitlement – but to receive the money, the parents must be registered as parents to the child in question and national identity numbers are required for the parents and child. This system covers all children and all parents. As the information in the system's registers is maintained and updated, all persons in the country will gradually be covered and the register will contain administrative, but also statistically important, links between all parents and children.

It is important for good coverage that the administrative systems cover both urban and rural populations, rich and poor citizens, and small and big enterprises. The ideal is that there is no selectivity. If suitable methods are not developed, selectivity will result in biased statistical estimates. For instance, in the Nordic countries all seriously ill persons will see a doctor, and all doctors know that cancer patients should be reported to the National Cancer Register. In this way we can be almost absolutely sure that all patients with a cancer diagnosis are in the Cancer Register. If rural or poor persons are underrepresented, estimated cancer incidence and mortality figures would be of low quality.

Unified systems of identity codes

Identities are important in administrative systems. Legally important relations between persons, such as husband and wife, or parents and children, are registered with the identities of the persons in question. In many registers the legally important relations between owners and different kinds of property are recorded with both the identities of owners and identity of property. For taxpayers, it is important that the tax paid is recorded together with the identity of the taxpayer. It is therefore in the interest of each taxpayer to use a correct identity in each transaction. The legal importance of identities explains why identity data as a rule are of high quality in many administrative sources.

The best way to handle identities in administrative systems is to use *national identity numbers*. Persons, enterprises and property should be given unique identity numbers that are used in all administrative systems in the country, and the same number should follow each person, enterprise or property over its lifetime.

Not only will administration become efficient; the statistical production system will become efficient when administrative data are used for statistical purposes, as it will be possible to link records and create important statistical comparisons. With unique national identity numbers, record linkage will be easy and the risk of false matches and false non-matches will be low. The statistical possibilities that national identity numbers create will be explained in the following chapters.

It is advantageous if the identity numbers have no relation to any attributes of the objects that are to be identified. For example, identity numbers for persons should not depend on name, sex, or address of the persons, because such attributes can change over time. Throughout the book we will use the abbreviation PIN for national identity numbers for persons and BIN for national identity numbers for legal units representing enterprises.

1.4.2 Legal base and public approval

There are preconditions concerning legal base and public approval that make possible the efficient use of administrative registers for statistics. These preconditions are discussed in UN/ECE (2007) and we build on that discussion here.

Legislation determines what data are generated

The national administrative systems for taxation and welfare are based on legislation that determines the kind of administrative data that are generated within these systems. If, for example, citizens pay income tax to municipalities, then the authorities must know where each citizen lives. The municipal taxation and welfare systems are the legal base for the Nordic administrative population registers. They are used not only for taxation and municipal welfare, but also for elections where the population register defines where each voter votes. For statistical purposes, this creates very good links between persons and geography that facilitate regional statistics. The administrative registers are updated every day, which makes possible timely monthly demographic statistics.

Legislation to improve the national statistical system

Politicians want to reduce the response burden of persons and enterprises as well as the direct costs for the production of community statistics.

- Legislation should provide the national statistical offices access to administrative microdata including identities, and the right to use the data for official statistics and research.
- Legislation should provide statistical offices the authority to match data from different sources and use data that were not originally generated for statistical purposes.

- Legislation could also instruct statistical offices to first use data from administrative registers and to conduct sample surveys or censuses only if available administrative data are insufficient.
- Some laws have the sole purpose of making register-based housing and population censuses possible. For example, the Nordic parliaments have decided that all employers must provide information on where all employees work – the local unit address for all. This information is given with income statements with data on employer identity, local unit identity, employee identity and wages and preliminary tax paid. These income statements play an important role in the Nordic statistical systems, as we obtain important links between three different object types. The parliaments have also decided that all persons should be registered at the dwelling where they live. It will then be possible to create statistics for households defined by the common dwelling in the register-based census.

Legislation on data protection

According to the second precondition in Chart 1.3, a national statistical office should have access to administrative registers kept by public authorities. This right should be supported by law and the protection of privacy must also be protected by law. Legislation that gives a statistical office access to administrative data is discussed above, and the protection of privacy and integrity are discussed below.

The principle of *one-way traffic* is important for data protection. Microdata can go from administrative authorities to the statistical office but never in the reverse direction.

The legislation on data protection should rest on a reasonable balance between protection of integrity on the one hand and increased costs and difficulties for statistics production on the other. An important task for top management at a national statistical office is to explain the consequences generated by proposed legislation to lawyers and politicians.

Public approval

The cooperation between register authorities and national statistical offices should be open and transparent. The fact that administrative data are used for statistical purposes should not be kept quiet; instead, the benefits and the efforts to protect integrity should be explained in open discussion and public debate.

It is important to explain that individual records regarding persons are anonymous in statistics production, in contrast to how administrative authorities handle the same data.

If the national statistical office has a good reputation as trustworthy, it will be easier to gain access to administrative data for statistics production. However, one mistake in the protection of integrity can immediately destroy this reputation.

Persons and enterprises do not want to be required to report to both an administrative authority and the national statistical office. Not having to do so will make public opinion more favourable to the use of administrative data for statistical purposes. It will become more difficult to motivate the double provision of data – why respond to a questionnaire on the enterprise's turnover when you also submit a value-added tax return to the Tax Agency which includes the same information?

Evidence that double provision of data to Statistics Sweden and to another authority is regarded as unreasonable can be seen in this newspaper clipping:

Translated from a newspaper article:

Refuse to send statistics to Statistics Sweden!

Mr R from the B-farm thinks that the authorities should be able to find the information from their own registers. Mr R refuses to send in statistics to Statistics Sweden. Because he already sends in information every other week to the Swedish Board of Agriculture, he thinks that the authorities should cooperate with each other instead. ...

1.5 Basic concepts and terms

Two principles form the basis of this book – the *survey approach* to administrative data and the *systems approach*. The survey approach means that we discuss estimates, estimators and quality as in a book on sample surveys. The systems approach builds on the *register system* concept that is introduced in Chapter 4 and is used throughout the book. We also discuss the *production system* at a national statistical office and the role of administrative registers in the design and development of that system.

We discuss three concepts in this section: what is a *statistical survey*, what is a *register* and what is a *register survey*? We also give examples of register surveys that illustrate some important principles discussed in later chapters: The Income and Taxation Register is a survey of persons and households and the Quarterly and Annual Pay Registers are business surveys.

1.5.1 What is a statistical survey?

This term is a central term used by statisticians at all national statistical offices. For many statisticians, however, the term is synonymous with *sample survey*. This will cause confusion when we discuss statistics based on administrative registers.

To avoid this confusion, we follow the distinction between different kinds of surveys that Statistics Canada (2009) use in their Quality Guidelines. The guidelines are written with censuses and sample surveys as the main focus. In this book, we focus on register surveys (3 below), but also discuss and compare other survey methodologies.

Statistics Canada, Quality Guidelines:

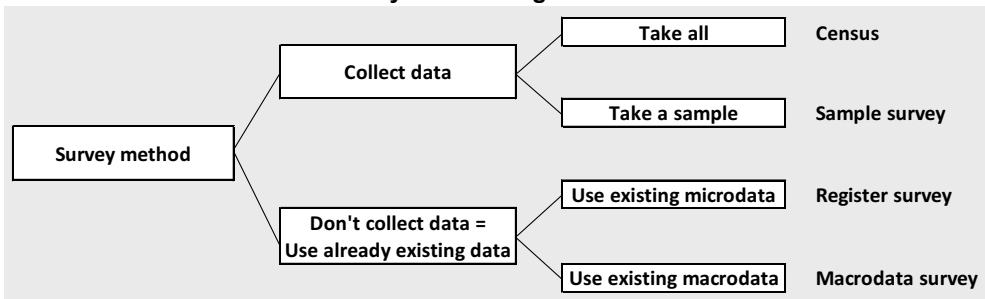
The term survey is used generically to cover any activity that collects or acquires statistical data. Included are:

1. a *census*, which attempts to collect data from all members of a population;
2. a *sample survey*, in which data are collected from a (usually random) sample of population members;
3. collection of data from *administrative records*, in which data are derived from records originally kept for non-statistical purposes;
4. a *derived statistical activity*, in which data are estimated, modelled, or otherwise derived from existing statistical data sources.

Estimates of, for example, number of employees by industry (as in Chart 1.2) can be based on a census, on a sample survey, or on a register survey. We can choose between these three different survey methodologies to estimate the same parameters. This is the reason why we have chosen to use the survey approach to administrative data – register surveys are only a new alternative to the two other well-established survey methods.

The fourth survey method above is the method that is used for the National Accounts. The National Accounts survey is based on a model-based compilation of macrodata (or estimates) from a system of economic surveys. Chart 1.4 compares the four kinds of surveys.

Chart 1.4 The four different survey methodologies



Sample surveys are based on a mathematical theory – probability and inference theory. Censuses and sample surveys are based on a non-mathematical survey methodology based on behavioural science – psychology and cognition are important aspects that are used to discuss errors that arise during the collection of statistical data through interviews and questionnaires.

Register surveys require a non-mathematical theory based on a systems approach. Macrodata surveys should also be based on a theory of systems of surveys. We discuss these issues later in this book when we introduce the concept of *survey system design*.

1.5.2 What is a register?

An administrative register is maintained to store records on all objects to be administered, and the administrative process requires that all objects can be identified. The following definition is valid for administrative and statistical registers:

A *register* aims to be a complete list of the objects in a specific group of objects or population. However, data on some objects can be missing due to quality deficiencies.

Data on an object's identity should be available so that the register can be updated and expanded with new variable values for each object.

– *Complete listing* and

– *known identities* are thus the characteristics of a register.

Catalogue, directory, list, register, registry are different terms for the same concept. We will only use the term *register*.

The following are examples of registers:

- Civic, civil or national registration of the population in a country results in registers of citizens, births and deaths.
- Income self-assessments from persons give registers of all taxpayers for a given year.
- In Sweden, enterprises with a turnover of SEK 40 million or more should report monthly. This gives monthly registers of all enterprises that have reported. For smaller enterprises, we obtain quarterly or yearly registers.
- All export and import transactions are registered by Customs. Monthly registers are created with all transactions for a specific month.
- A census file with data from a housing and population census is a register if there are identities of the persons in the file.

The identities used in register processing can either be identity numbers that are unique within a national administrative system or an identity number in a subsystem with keys to the identities in other systems. It is also possible to use identities defined by, for instance name, address, date of birth and place of birth.

These identities will be used in *deterministic matching* of the objects in different registers, where the aim is to find identical or related objects in two registers. In deterministic matching, two records are linked if the identifiers agree exactly. This is the most efficient method when the identifying variables are of good quality.

Chart 1.5 Deterministic matching with Personal Identity Numbers, PIN

Population register			Administrative income register		Combined register after exact matching				Statistical income register after imputations			
Person	Sex	Age	Person	Income	Person	Sex	Age	Income	Person	Sex	Age	Income
PIN1	F	87	PIN1	167	PIN1	F	87	167	PIN1	F	87	167
PIN2	M	74	PIN2	215	PIN2	M	74	215	PIN2	M	74	215
PIN4	M	62	PIN3	94	PIN3	*	*	94	PIN3	*	*	94
PIN5	F	49	PIN4	341	PIN4	M	62	341	PIN4	M	62	341
PIN6	F	35	PIN5	298	PIN5	F	49	298	PIN5	F	49	298
PIN8	M	14	PIN6	277	PIN6	F	35	277	PIN6	F	35	277
					PIN8	M	14	*	PIN8	M	14	0

Because person PIN3 is not in the population register and person PIN8 is not in the administrative income register, the combined register after deterministic matching will have two records with missing values due to this non-match.

Many administrative registers consist only of persons or enterprises of a defined category. Only persons with income are in the administrative income register in the example in Chart 1.5. When such registers are combined with the population register, the non-match will generate missing values. Zero income must be imputed for persons not in the administrative income register, such as person PIN8. Person PIN3 is not in the population register and if that person is not found in any other register the non-match will result in missing values (*) for sex and age.

1.5.3 What is a register survey?

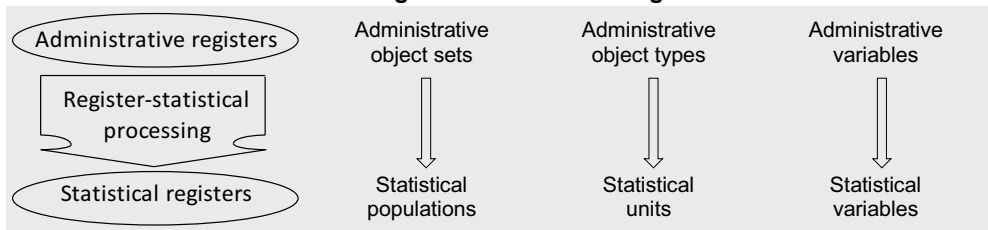
The original data are generated in public administrative systems. Definitions of object sets, objects and variables are adapted to administrative purposes. Every authority carries out controls, corrections and other processing suited to their administrative aims.

When an authority delivers data to a national statistical office, further selections and processing may be carried out to meet the needs of the statistical office. The authorities also have metadata as definitions, administrative rules and quality aspects, based on the administrative authority’s experiences and investigations. This information is important for those receiving the data at the statistical office.

It is generally not a good idea to produce statistics directly from the received administrative registers because these are not adapted to statistical requirements. The object sets, object definitions and variables need to be edited, and as a rule it will be necessary to carry out some processing so that the register fulfils the statistical requirements for population, objects and variables. The register-statistical processing, which aims to transform one or several administrative registers into one statistical register, should be based on generally accepted statistical methods.


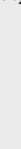

Chart 1.6a shows three important components of this work. We have found that people have a tendency to use administrative concepts as they are, and in some cases this can be acceptable – but in other cases it can be unacceptable. The three issues of how to define population, units and variables of a statistical register are important for the quality of the statistics to be produced with the newly created statistical register.

Chart 1.6a From administrative registers to statistical registers



A statistical population or administrative object set consists of N objects or units or elements. Of these three synonyms, we will as a rule use the term *object* for the units in an administrative object set and the term *statistical unit* for the units in a statistical population. The register-statistical processing is described in Chart 1.6b.

Chart 1.6b From administrative registers to statistical registers

Administrative object sets  Statistical populations	Matching object sets Handling of non-match Selection of objects Processing of time references	Administrative object types  Statistical units	Editing to find errors in objects and false matches Handling of missing objects Creating derived objects	Administrative variables  Statistical variables	Editing to find wrong variable values Handling of missing values Coding Creating derived variables
--	--	---	--	--	---

1.5.4 The Income and Taxation Register

The Income and Taxation Register (I&T) is an important part of Statistics Sweden's register system. It is used to describe income distribution and for regional income statistics, and it is the basis for longitudinal income registers used by university researchers.

This register utilises many administrative sources, and many administrative variables are used to create important statistical variables. Besides these administrative sources, it is necessary to use the register system at Statistics Sweden: the Population Register is used to define the population of the Income and Taxation Register, and important classification variables are imported from other registers in the system to the Income and Taxation Register.

1. *Data generation at the National Tax Agency*

The annual income self-assessment is based on tax returns from income earners and the taxation decisions of the local tax authority. Both the income earner and the tax authority use statements of earnings for salary, sickness benefits and interest payments that are the responsibility of employers, social insurance office and finance companies. The National Tax Agency ultimately compiles this information. Tax returns, statements of earnings and taxation decisions can be changed and supplemented. Data for one person can thus be very complex.

2. *Microdata deliveries to the Income and Taxation Register*

The Swedish National Tax Agency annually creates databases that contain information on Sweden's population. The data files for one year – containing around 9 million records, each with around 300 variables – are delivered to the Income and Taxation Register at Statistics Sweden.

3. *Metadata to the Income and Taxation Register*

Record descriptions with names and definitions of variables accompany the deliveries from the National Tax Agency. Tax return forms, statement of earnings forms, taxation decisions and tax return instructions are also necessary for the correct interpretation of the data.

4. *Editing of data*

The I&T Register receives data from many different suppliers outside and inside Statistics Sweden. External data are edited. Data from other Statistics Sweden registers have already been edited. Contacts with suppliers are important to obtain knowledge of changes in the administrative system, which in turn is important to ensure the quality of the register statistics – administrative changes should not be interpreted as actual income changes.

5. *Matching and selections*

There is a large number of registers that should be processed to create the different sub-registers that are included in the Income and Taxation Register. Records from different sources are matched using Personal Identification Numbers (PIN), and aggregation is carried out at the same time, i.e. all the statements of earnings data for a specific person are aggregated so that the person's income from work can be put together. One type of processing is to select persons aged 16 and older who were also parts of the population on 31 December.

6. *Derived objects are created*

More information on certain relations helps to form household units. Between adults, the relations *married* or *cohabiting adults with children in common* result in their placement in the same family unit. These relations are derived from the family members' personal identification numbers; these reference variables are found in the taxation data and in Statistics Sweden's Population Register.

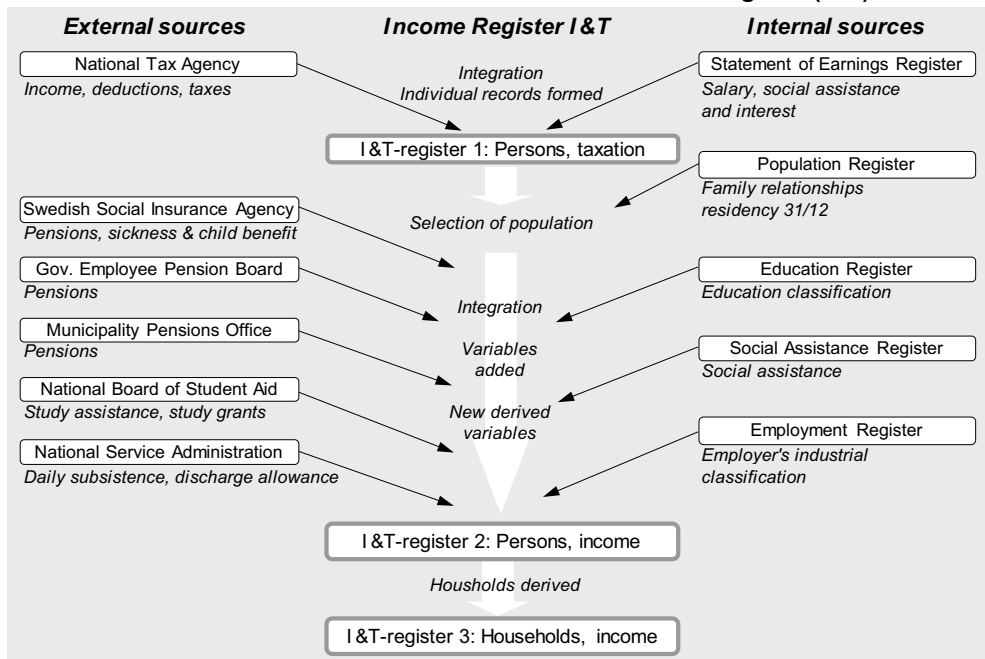
7. *Derived variables are created*

A large number of derived income variables are formed. For instance, the wage or salary amounts are aggregated from the different earnings data to become an individual's *income from work*. Every person's total income from work and capital plus transfer payments minus tax becomes the person's *disposable income*. For households, variables such as *household type*, *number of consumption units* and *disposable income* are formed.

Chart 1.7 shows how the Income and Taxation Register receives administrative data from a variety of different external sources and some Statistics Sweden registers. The middle column shows the different phases when the sources are used during the process to create the new statistical register.

This example shows the importance of the principles in Chart 1.1. Statistics Sweden has access to many administrative registers with variables describing different kinds of income. The object set and the administrative variables have been processed to meet statistical needs. Many sources have been used to produce a statistical income register with rich content. The population in the income register is consistent with other statistical registers within the register system.

Chart 1.7 Different data sources for the Income and Taxation Register (I&T)



1.5.5 The Quarterly and Annual Pay Registers

Aggregate wages and salaries by economic activity and institutional sector are important inputs for yearly and quarterly National Accounts. One quarterly and one annual register survey produce these estimates and we will use these surveys to illustrate the different ways a source can be used as noted in Section 1.3.

A simple and straightforward approach

An administrative register of good quality can be used almost as if it were for statistical purposes. If content and coverage are sufficient, it will be relatively easy to use the administrative register and produce statistics of good quality. The different steps of the yearly survey are illustrated in Charts 1.8a–1.8c.

The administrative Annual Pay Register is first *edited* to find incorrect or unreasonable values of aggregated wages and salaries (WagesYear). The identity number (BIN) is first checked to ensure that all values have the right format and acceptable values. Preliminary tax (Prel-tax) should be between 30% and 35% of WagesYear and this relation is used to edit WagesYear. For the enterprise with identity number BIN05, we find a 1000-factor error; 2 is therefore replaced with 2 000 and the variable W-imp is created to show imputed values of WagesYear (Chart 1.8b).

The registers in Chart 1.8a are then *matched* with the identity numbers BIN. Records with the same value of BIN are combined in a new register that is shown in Chart 1.8b.

Chart 1.8a The sources of the Annual Pay Register

1. Business Register			2. Administrative Annual Pay Register				
BIN	Sector	ISIC	BIN	WagesYear	Source	Prel-tax	
BIN02	6	52	BIN01	25	I	8	Sector: 1 Non-financial enterprises 2 Financial enterprises 3 Government 4 Municipalities 6 Self-employed 7 Non-profit organisations
BIN03	1	51	BIN03	1 667	I	544	
BIN04	7	91	BIN04	796	I	252	
BIN05	1	70	BIN05	2	P	689	
BIN06	1	45	BIN06	92	I	29	
BIN07	1	51	BIN07	4 758	I	1 565	
BIN08	1	60	BIN08	39	P	12	
BIN09	1	28	BIN09	452	I	142	
BIN10	1	74	BIN11	289	P	95	
BIN11	1	27	...				
...			Count	305 411			
Count	331 518						

When the 331 518 records with active employers in the Business Register are matched with the 305 411 records in the administrative Annual Pay Register we find that there is a lot of *non-match*. There are 33 543 records in the Administrative Register missing in the Business Register; this is an indication of *undercoverage*. And 59 650 records are missing in the administrative register; this is an indicator of *overcoverage* in the Business Register. The non-match is shown to the right of Chart 1.8b below.

This non-match above gives rise to *missing values* (*) in the variables Sector and Economic Activity, ISIC. After imputations the final statistical Annual Pay Register in Chart 1.8c can be created.