Shifting paradigms in official statistics

From design-based to model-based to algorithmic inference

Bart Buelens, Harm Jan Boonstra, Jan van den Brakel and Piet Daas

Discussion paper (201218)

Statistics Netherlands

The Hague/Heerlen, 2012

Explanation of symbols

.	data not available
*	provisional figure
**	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/'12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/'10– 2011/'12	crop year, financial year, etc. 2009/'10 to 2011/'12 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201218 X-10

SHIFTING PARADIGMS IN OFFICIAL STATISTICS
FROM DESIGN-BASED TO MODEL-BASED TO ALGORITHMIC INFERENCE

Summary: Official statistics are aimed at providing reliable estimates of population quantities. Inference is traditionally motivated from a design-based perspective, with the model-based approach being gradually adopted in specific circumstances. We take this shifting paradigm one step further, from model-based to algorithmic inference methods.

Surveying a sample of the population of interest – typically enterprises or households – is fundamental to the design-based approach, where the design is the basis for inference. Model-based estimation methods may provide a viable alternative in situations where design information is not available. Estimation of the model parameters is pivotal, although in official statistics it is only an intermediate goal, as the model is ultimately used for prediction. Therefore, adopting a data-centred, algorithmic view rather than a model-centred view is possible. The algorithmic view encompasses methods generally attributed to the fields of data mining, machine learning, or statistical learning. Algorithmic methods may be useful in situations where data are not obtained through a sample survey, and where the typical models used in model-based estimation are not tenable.

These different methods of inference are discussed, and illustrated through a simulation study. Some preliminary experiments conducted at Statistics Netherlands are presented, using social network and mobile phone data.

Keywords: data mining, statistical learning, official statistics

1. Introduction

Sample surveys are expensive and cause response burden. National Statistical Institutes (NSIs) are under increasing pressure to use alternative sources of data, such as registers and administrative data, or even other sources. These sources are usually collected or maintained by government agencies or private companies. Sometimes data that were never intended for statistical use are explored.

The classical tool set of the practitioner of official statistics is not always capable of handling these new types of data. In particular, the issue of deriving estimates of population quantities from such data sets can be a daunting task.

The present paper formulates a framework in which new methods are positioned next to well accepted methods typically used at NSIs. It motivates research into methods of inference capable of handling data sources currently not used for official statistics production yet, but likely to become an integral part of NSI production processes in coming years.

Section 2 discusses types of data sources used at NSIs. Section 3 presents three classes of methods of inference, and explains how these relate to the types of data sources. A simulation study illustrating the methods is presented in section 4, and two real-world examples in section 5. Section 6 concludes this paper with suggested directions for further research.

2. Data for official statistics

2.1 Data types

Many statistics produced by NSIs are based on sample surveys. The data collected through a survey are referred to as primary data. In a survey, the values of a variable of interest are observed for part of the population. A sample data file contains records corresponding to units of the population of interest, with each record containing a number of variables of interest – target variables, survey variables – and possibly some known characteristics of the units. An example is the Structural Business Survey, where the units are enterprises and an important survey variable is turnover.

Secondary data files are similar in structure, but are not the result of a sample survey. They are typically collected in support of some administrative process. The variables are not chosen or defined by the NSI, as opposed to the variables in primary data files. Examples of secondary data are the population register, and VAT-data. The fundamental difference with sample data is the manner in which secondary data come about.

We believe that a third type of data can be distinguished, namely data that are not comprised of a set of records directly corresponding to units in a target population. These kinds of data sources often register *events*. In the present paper the term *tertiary data* will be used for this kind of data sources. Such data can be generated as a by-product of some process unrelated to statistics or administration. Since these data files are often much larger in size than sample data or administrative registers, the term 'big data' is sometimes used in these cases. Leading scientific journals have recently dedicated special issues and online content to this topic (Nature, 2008; Science, 2011). Examples of tertiary data are all the mobile phone calls made in a certain period of time including their originating location, or all social media messages created within a certain period of time in a specific region.

The terminology presented here is debatable, and different definitions and terms will be found in the literature. Table 1 gives an overview of the characteristics of primary, secondary and tertiary data as used in this paper.

An essential ingredient to enable inference for unobserved population units is the presence of auxiliary data, or covariates, which are characteristics known for all units of the population of interest, including those units for which no target variables are available. If no auxiliary data are available, one must ascertain that the data set

covers the complete population, or assume that the data set is representative of the population. The problem of missing covariates can sometimes be dealt with through data integration. Data integration is a powerful approach to link different data files (Bakker, 2011), thereby potentially combining survey data with administrative registers or big data sources; the latter after conversion of course.

The most important difference between tertiary data and the other two types is the structure and the availability of target variables. In the case of tertiary data, dedicated preprocessing or data preparation is required to transform the data to units of interest, and to derive the target variables.

Another difference between these data sets is related to the measurement errors they can contain. Survey data may contain measurement errors in the sense that respondents provide a wrong answer, either deliberately or not. Registers are subject to a range of potential errors, for example records that are not up-to-date (Wallgren and Wallgren, 2007; Zhang, 2012a). The errors found in tertiary data are sometimes of a different nature, in the sense that, rather than human or administrative errors, the errors are often due to machine or system errors or failures. Examples are power black-outs and ill-calibrated sensors.

Table 1: Main characteristics of the three data types.

	primary	secondary	tertiary
records are units of target population	yes	yes	no
target variables directly available	yes	yes	no
auxiliary variables directly available	yes	often	no
data preparation/conversion needed	no	no	yes
data covers complete target population	no	often	rarely
data are (almost) representative	usually	usually	no
susceptibility to measurement error	high	medium	low

2.2 Data preparation

Data from sample surveys or administrative registers is typically ready for use in inference and for analysis. The records in the data files correspond to units in a target population, and the variables of interest are available within those records.

With tertiary data, this is usually not the case. The processing needed to convert such data sets into the required form may be non-trivial, and is often referred to as data preparation, preprocessing, or feature extraction (Duda et al., 2001; Pyle, 2009). The task is to transform the available event-based data to unit-level data – where the units are the statistical units of interest – together with some associated variables. While tertiary data typically contain fewer measurement errors, this additional processing may come with its own types of error.

Preprocessing methods are seldom generic. They are often suited only to a particular kind of data, e.g. images, text, sound, etc. See section 5 for two examples.

Sometimes in the literature, preprocessing in itself is referred to as data mining (Duda et al., 2001). But in general, more is needed in order to produce official statistics, in particular in cases where the data set does not cover the complete target population.

In official statistics, interest is say in a total Y of some variable y for some target population. If not all y values are known for the complete population, some method of inference is needed to estimate the missing y values, or at least their aggregated contribution to the total. In addition to the variable of interest y, some other characteristics of the units of the population are known. These auxiliary data are denoted with x. For survey and register data, y is available, and x is often easily obtained through linking with population registers. Tertiary data is very different in this sense, as it does not contain y values directly. The derivation of these values may be non-trivial, specific to a particular data type, and therefore not generic. Auxiliary data is also likely not or only indirectly available, but could potentially be linked after conversion.

In the present overview it is assumed sensible y values can be obtained from event-based data files, and data integration techniques allow for extending the data sets with covariates x.

In many applications seen in the literature or on the internet, x is unavailable, and it is assumed that the data set covers the complete population, in which case simply aggregating over all available data is perceived sufficient in order to estimate Y. For commercial applications this may be worthwhile and informative. For example a bank can make statistics about energy bills of their customers, since these bills are generally paid by bank transfer, and the transfers are easily recognized. Similarly, an online bookstore can make statistics about their top selling titles. Important in the context of official statistics is that in these examples the statistics apply to customers of that bank and of that bookstore respectively, and cannot be directly generalised to the whole population.

The problem with this straightforward approach is that the available data may be a selective subset of the target population; a subset not representative of the target population. Assuming the statistics are valid in spite of the lack of representativity is not acceptable (Kruskal and Mosteller, 1980), as it may lead to biased results. In the following section methods are listed that enable generalizing outcomes from a subset of the population to the whole population: inference.

3. Inference in official statistics

3.1 Background

How to estimate population quantities such as totals and means when data are not available for all units in the population? Methodologies of inference present answers to this problem. This section gives an overview of methods, discusses the relation with the previous section, and highlights some additional aspects including selectivity, and inference for completely observed populations.

3.2 Methods of inference

3.2.1 Design based inference

In official statistics, the typical approach to obtain estimates of population quantities is conducting a sample survey. A target variable *y* is observed for a limited sample of units from the population. The selection of the sample is randomized, according to some design. Stratified, multi-stage and cluster sampling are commonly used sampling designs. Based on a design, population units *i* have probabilities π_i of being selected in the sample. Such man-made randomization is a vital prerequisite.

Inference from the obtained sample proceeds using the observed values of the target variable y_i, and the design information which is available through the inclusion probabilities. An estimator of the population total *Y* is

$$\hat{Y}_D = \sum_{i \in S} \frac{1}{\pi_i} y_i \qquad (1)$$

with *S* the sample. This basic design based estimator is known as the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). Enhanced versions improve the precision of this estimator and attempt to correct for selective non-response by including auxiliary information in the estimation procedure. Sarndal et al. (1992) is a classic reference in this context.

An important and in official statistics highly appreciated property of the estimator is that it is design unbiased, $E(\hat{Y}_D) = Y$. This means that the estimate averaged over all possible samples drawn according to the design, is equal to the population quantity *Y*. Any given sample may of course give rise to an estimate which is not equal to *Y*.

The uncertainty associated with the estimate $\hat{Y}_D$ arises from the incomplete observation of the population. If the complete population was observed, all $\pi_i = 1$ and $\hat{Y}_D = Y$. The larger the sample, the smaller the uncertainty associated with the estimate.

Design-based estimation does not only use auxiliary background variables in the estimator, but also in the sampling design. For example, in stratified sampling where

inclusion probabilities vary between strata, the stratification variable is an example of an auxiliary variable used in the design, and hence also in the estimator.

3.2.2 Model based inference

When auxiliary variables *x* are available that correlate with the target variable *y*, this correlation can be exploited through explicit modeling (Van den Brakel and Bethlehem, 2008; Vaillant et al., 2000). The most elementary model is a linear model,

$$y = \beta_0 + \beta x + \varepsilon \tag{2}$$

expressing *y* as a linear function of *x*, with ε an error accounting for the linear relationship being not perfect. Typically, *x* is a vector.

In situations where variables *x* are available for the whole population, and *y* only for a limited subset, model based inference can be applied. It encompasses fitting a model using the units for which both *x* and *y* are available, and using the fitted model to predict the unknown *y* values using the *x* as predictors. Fitting a model means estimating the model parameters. With these estimates, $\hat{\beta}_0$ and $\hat{\beta}$, the model based estimate of the population quantity Y is

$$\hat{Y}_M = \sum_{i \in S} y_i + \sum_{i \in R} (\hat{\beta}_0 + \hat{\beta} x_i) \tag{3}$$

with *S* the set of units for which *y* is known, and *R* the complement of *S*. The unknown *y* values of units in *R* are predicted using the fitted model. The uncertainty in the model based estimate arises from the model not fitting the data perfectly. Comparisons of model-based and design-based estimation are found in Little (2004) and Gelman (2007). This debate, in particular the role of modelling, has a long history (Bethlehem, 1983) and is still ongoing (Van den Brakel, 2012).

An alternative to model (2) is to model the propensity for units to be present in the data set

$$\rho = f(\beta_0 + \beta x + \varepsilon). \tag{4}$$

This approach is largely inspired on classic approaches to non-response modelling in sample survey settings, where the response propensity is modelled according to a model similar to model (4). Typically, logistic regression models are used as opposed to linear ones, as the independent variable is binary. The function *f* in expression (4) is then the logit function.

Estimating model (4) results in estimated propensities $\hat{\rho}_i$ for all units in the population. A propensity score estimator is given by

$$\hat{Y}_P = \sum_{i \in S} \frac{1}{\hat{\rho}_i} y_i \,. \tag{5}$$

Note that $\hat{\rho}_i$ does not depend on y, which is an attractive property in case there are many target variables. On the other hand, correlations between x and y are not exploited. Although conceptually different, this estimator bears resemblance to the design-based Horvitz-Thompson estimator given in formula (1).

3.2.3 Algorithmic inference

Viewing modelling from a different perspective (Breiman, 2001), a model can be considered as a function F mapping units from their known values x to their unknown y, $F(x) = y$. This prediction view of modelling allows for extending the class of functions from which F is chosen. In principle, F can be any algorithm capable of mapping inputs x to outputs y. In the algorithmic approach, the equivalent of fitting a model is tuning an algorithm, so that it predicts well. If this does not involve the estimation of some parameter of a function, the term non-parametric method is used. It is generally impossible to express algorithmic methods analytically in terms of a mathematical model. This sets them apart from the model based methods.

In the prediction approach, the data for which both x and y are known is split into two parts. One part is used to tune the algorithm. This is often referred to as learning, or training. The second data set is used to evaluate – or test – the predictive capabilities of the trained algorithm.

Similar to the model based estimator, the algorithmic estimator is

$$\hat{Y}_A = \sum_{i \in S} y_i + \sum_{i \in R} F(x_i)\,. \tag{6}$$

with S the set of units for which y is known. It is this data set that gets split into two parts in the training stage. The set R contains the population units with unknown y.

Uncertainty of this estimator arises from the imperfect predictive power of the algorithm, and is assessed on the test set using some cost function.

Two algorithms discussed in Breiman (2001) are used in the simulation study in section 4, Classification and Regression Trees (CART), and Artificial Neural Networks (ANN). Details and additional references for these methods can be found in Breiman (2001) and Hastie et al. (2003). The methods are briefly summarized as follows.

CART builds decision trees based on a number of rules that are learned from the data. With binary trees – which are used in this paper – the value that best differentiates two groups in the data set, based on one variable, is determined. The data is split according to this value, forming two *branches*. Each branch is split again following the same procedure, until groups are sufficiently homogeneous.

An ANN is a network of basis functions that allow for the approximation of virtually any non-linear function. Networks with one layer can be seen as traditional functional decompositions. Networks consisting of multiple layers are sequences of such decompositions, allowing for even more flexibility. Various network layouts

and methods of training the networks exist. In this paper a so-called feed-forward network with one hidden layer consisting of five nodes is used.

3.3 Inference and data types

Model based and algorithmic inference can be jointly referred to as predictive inference, or statistical learning (Hastie et al., 2003). This requires the modeller to adopt the predictive view of modelling, rather than the explanatory one. The latter perspective is often taken when seeking to explain an outcome y by a set of potential auxiliaries x, and studying and identifying the most important ones. In inference, the goal is prediction of unseen y values given x, hence the prediction approach can be taken.

A distinctive property of inference for official statistics, compared to some other applications in data mining, is that the unit level predictions of y are of no immediate interest on their own. It is the population quantity obtained from the unit values that is of interest. For example, a prediction algorithm capable of classifying e-mails into "good" or "spam" is in typical data mining settings used to filter incoming e-mails. In the context of inference, it would be used to obtain the proportion of spam e-mails received in a 24 hour period, say, without necessarily studying unit-level predictions about individual e-mails.

At the risk of simplifying matters, Table 2 provides a framework in which data types and inference methods are listed together. Design-based methods are only applicable when there is a known design underpinning the data. When design information is lacking, the model based propensity estimator (5) somewhat resembles a design-based estimator, and may be attractive when many target variables must be estimated from the same data set.

While predictive methods can be applied to survey data, there is no pressing need to do so. Design-based estimators known to perform well need not be abandoned. In some situations modelling can be helpful; in sample survey settings, for example when sample sizes are small (Rao, 2003) or when the levels of non-response are high (Van de Brakel and Bethlehem, 2008).

In a statistical context, secondary data refers to administrative registers, which are often complete, or nearly so, or cover at least such a large proportion of the population that they can sensibly be assumed to be representative of the full population. Estimation of the missing parts of registers can be conducted using predictive methods. When the set of units for which y is unknown is much smaller than the set where y is known, the actual method of inference will have limited impact on the estimate of the population quantity.

The situation is different for tertiary data, which is characterized by being selective and having a rather limited set of auxiliaries x, a consequence of the fact that it is event based. In this case it is essential to determine the best method for inference, since it may crucially affect the population estimate. The scope must not be limited

to (generalised) linear modelling, but must include non-parametric or algorithmic approaches.

Table 2. Combining data types and methods of inference.

	Data type		
Inference method	**Primary**	**Secondary**	**Tertiary**
Design-based	Common in official statistics: survey sampling and design based estimation.	As there is no design, design-based estimation is not applicable	
Predictive	Predictive inference can be applied in principle, but is generally not necessary.	As registers are often (nearly) complete, fairly simple modelling suffices in most cases.	Tertiary data is often highly selective; consideration of algorithmic methods is beneficial.

3.4 Completely observed populations

In the preceding sections the assumption is made that the variable of interest is not available for the complete population. However, some registers or tertiary data sets may be complete. In this situation, with y known for all population units, the population quantity does not need to be estimated, but is directly calculated from the data. In the case of a total, this is simply an aggregation. This sum has no associated uncertainty, provided that there are no measurement errors.

Nevertheless, analysis aimed at, for example, comparisons between subpopulations, or monitoring change over time, may lead to peculiar situations. Since there is no uncertainty in the "estimate", any observed difference or change will be statistically significant, in the sense that it cannot be caused by some random element in the data collection or inference stages.

For example, if in a given year the percentage of people receiving unemployment benefit is 5.500%, and the following year it is 5.501%, there is an increase. It is statistically significant – since there is no uncertainty – but is it (statistically) relevant, in the sense of being an important socio-economic fact which should be taken into account in labour market policy making? Similarly, the percentage of recipients in two villages can be compared, or even in two neighbourhoods. But if the neighbourhoods are very small, say 100 people each, how meaningful is a difference of 1 percentage point, which is equivalent to one person?

The same questions arise when a very large proportion of the population is sampled. Small differences become significant with increasing sample sizes. Observing the complete population can be seen as the limit of increasing the sample size to its maximum.

One approach in such situations is to shift interest to specific parameters of a super population model, of which the finite population is considered one specific realization, see e.g. Valliant et al. (2000). In contrast to the finite population quantities, the model parameters will never be without uncertainty.

For dealing with temporal change, a time series model can be assumed. For cross sectional comparisons involving small subpopulations, ideas may be borrowed from the field of small area estimation (Zhang, 2012b).

These issues of statistical significance versus statistical relevance may be new to the official statistics practitioner, who is typically accustomed to dealing with samples that are small compared to the population size. The present paper does not address this issue further.

3.5 The selectivity issue

The fundamental aim of all the methods of inference presented above is in fact removal of selectivity of the data set used for inference. Primary data may not be representative because of selective non-response (Bethlehem et al., 2011). Secondary and tertiary data may be selective due to the process by which they are collected or generated. All estimators presented here assume that selectivity with respect to the target variables y can be explained and hence corrected through a set of auxiliaries x.

However, it is possible that x is not sufficient in explaining selectivity, as selectivity may depend to some extent on y itself. In sample survey settings, *not missing at random* (NMAR) non-response occurs when non-response behaviour is correlated with y, after controlling for x. Sample surveys are said to have *informative designs*, when the selection probabilities of the sample units are correlated with y. Pfeffermann (2011) discusses these issues in detail. Given a sample which is selective with respect to y, inference procedures ignoring this may lead to biased results, referred to as selection bias. One approach to dealing with selection bias was proposed by Heckman (1979). As such methods strongly rely on distributional assumptions, alternative approaches in the form of sensitivity studies have been proposed (Andridge and Little, 2011).

Such ideas could be adopted not only for primary data but more generally, also for secondary and tertiary data. This is an area of research worth investigating further. In the present paper only situations where selectivity is properly explained by the set of auxiliaries are considered, as is generally done in mainstream survey sampling text books (Sarndal et al., 1992; Bethlehem et al., 2011).

4. Simulation study

The methods discussed in section 3.2 are applied in a simulation study. The study is based around a publicly available data set from the UCI repository (Frank and Asuncion, 2010). It contains physical measurements of abalone – an edible sea snail common in e.g. Australia – and their age (Nash et al., 1994). For the present simulation, the variables rings (=age), length and (whole) weight are used.

A population of size 100,000 is created through repeated sampling from the original 4,177 records. The correlation in the original file between length and weight is 0.93. This is reduced to 0.85 in the simulation, by adding a random normal disturbance to the length variable in the artificial population. Weight is used to create a categorical variable, identifying abalone as light if their weight is less than 100 grams, and heavy otherwise. Approximately 30% of the population is light according to this definition.

A data set to be used for inference is obtained through stratified sampling from this artificial population. The newly derived weight class is used as a stratification variable, with a sample of total size 2,000 allocated equally to both strata. Consequently, individuals in the light weight stratum have higher inclusion probabilities than in the heavy weight stratum.

The three methods of inference are applied to this sample, each with some different variants. The following estimators for the average age of the abalone in the population are compared:

- dsgnHT Horvitz-Thompson estimator,
- dsgnSM Horvitz-Thompson ignoring the design (= sample mean),
- predM1 model-based, linear model incl. length and weight,
- predM2 model-based, linear model incl. weight class,
- predM3 model-based, linear model incl. inclusion probability,
- predN1 ANN with length and weight as input,
- predN2 ANN with weight class as input,
- predT1 CART with length and weight as input,
- predT2 CART with weight class as input.

The ANN methods use neural networks with one hidden layer and 5 and 3 hidden neurons in predN1 and predN2 respectively. The CART methods implement binary trees.

To avoid peculiarities of an individual sample, and to get an impression of variance, this procedure is bootstrapped 1,000 times, and the results are averaged. The bootstrap mean and variance are used to compare the methods, see Figure 1.

The Horvitz-Thompson estimator is design unbiased when inclusion probabilities are taken into account (dsgnHT). Ignoring the design, this estimator reduces to the

sample mean (dsgnSM), which is biased due to the overrepresentation of lightweight abalone. The model based estimator using the continuous weight and length variables, predM1, is biased low. This can be explained by the fact that the relation between weight and age is not linear, see Fig. 2. A linear model is obviously not appropriate in this case. When the weight class is used, as opposed to weight as a continuous variable (predM2), the bias disappears. This is exactly the same as including the inclusion probabilities as covariates in the linear model (predM3), as the two categories of weight class span the same linear space as the inclusion probabilities plus the constant term. These models no longer model the relation between weight and age, but rather the age difference between the two weight classes, which explains the fact that they are not biased. The three estimators

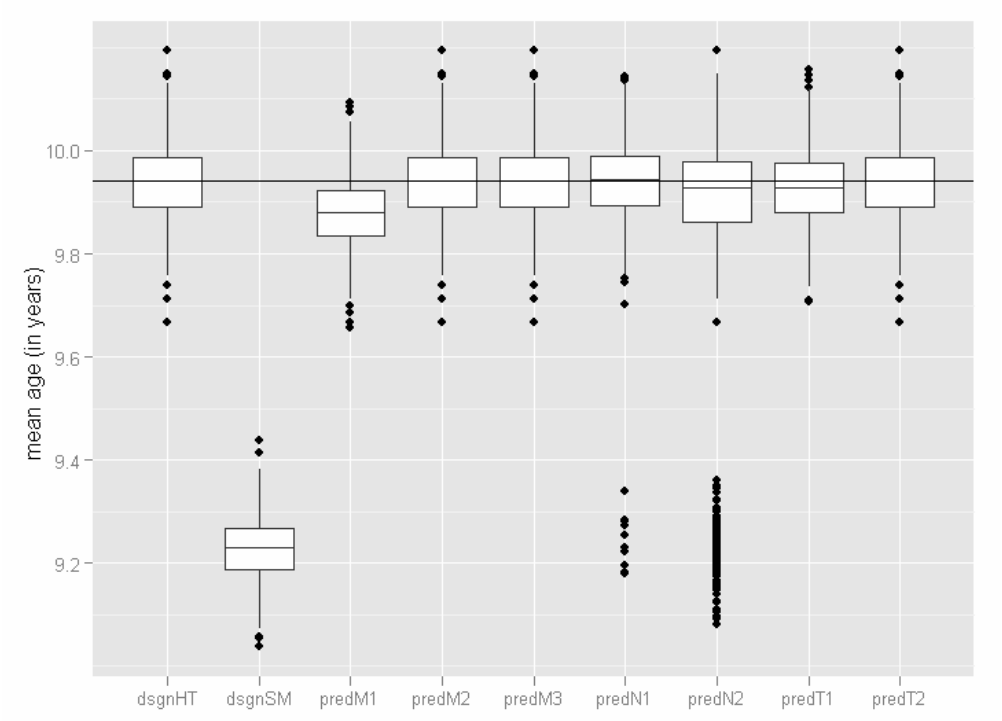

Fig. 1. Results of the simulation study. For each estimator the boxplot represents the distribution of the 1,000 bootstrap estimates. The horizontal line indicates the known population mean.

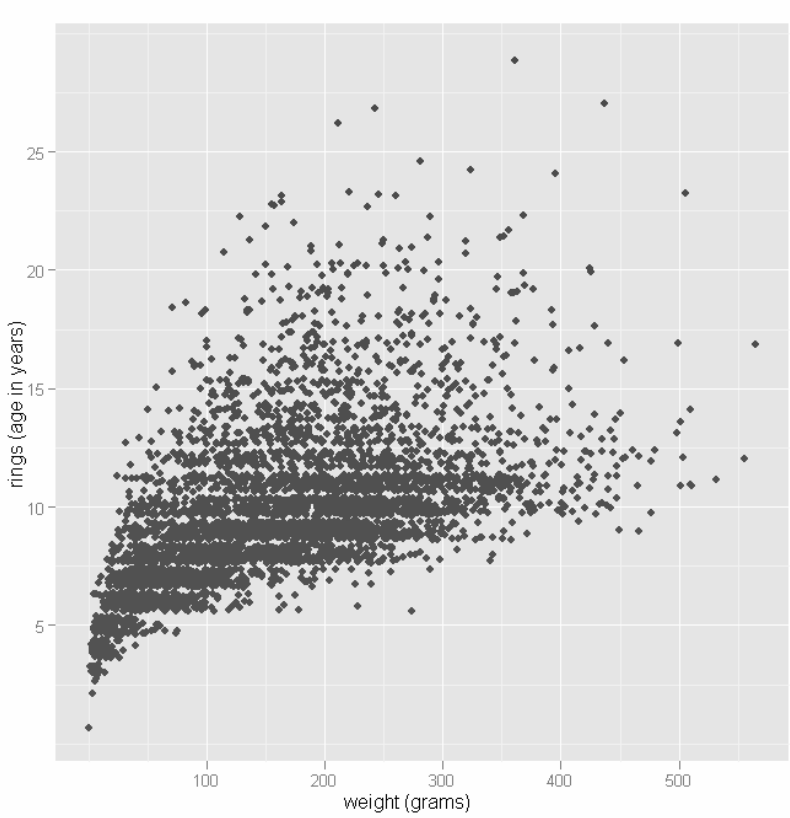

Fig. 2. Age versus weight in the artificial abalone population (N=100,000).

dsgnHT, predM2 and predM3 use exactly the same information and indeed give almost identical results.

The neural network prediction predN1, using continuous weight and length as inputs, is unbiased. Contrary to the linear model estimator predM1, this estimator is capable of handling the non-linear relation between the inputs and the output. The predN2 estimator which only uses the weight class as an input is somewhat biased, and very unstable. The variance is rather large. This seems to be caused by instances where the ANN is unable to distinguish the two classes, as there is a group of bootstrap results concentrated around the same level as the sample mean, dsgnSM. This is the case too for PredN1, but to a lesser extent.

The regression tree estimator predT1 is slightly biased low, probably for similar reasons as the linear model based estimator predM1. The bias disappears when using the weight class as input (predM2).

Bias, standard deviation and mean square error (MSE) of all estimators are given in Table 3. In MSE-terms, dsgnHT, predM2, predM3, predT1 and predT2 perform equally well. All these estimators use the same auxiliary information, either directly through the weight class variable, or indirectly through the design weights (inverse inclusion probabilities), with the exception of predT1. The latter estimator is the best among all estimators using both continuous weight and length variables, although it is more biased than the predN1 estimator.

Table 3. Bias, standard deviation and MSE of the estimators in the simulation study.

	Bias	*St. dev.*	*MSE*
dsgnHT	-0.001	0.071	0.0051
dsgnSM	-0.713	0.060	0.5118
predM1	-0.061	0.066	0.0081
predM2	-0.001	0.072	0.0051
predM3	-0.001	0.072	0.0051
predN1	-0.006	0.100	0.0099
predN2	-0.089	0.246	0.0683
predT1	-0.013	0.070	0.0051
predT2	-0.001	0.072	0.0051

This simulation illustrates the power of design based estimation in survey sampling. At the same time it illustrates the power of the predictive methods, which perform fairly well even without design information. Success of these other methods crucially depends on the availability of a covariate which explains selectivity or skewness in the data set, in this case the over representation of light weight abalone.

Many surveys conducted at NSIs suffer from decreasing response rates. Non-response is often selective. Methods correcting for selective non-response are available, but are critically reliant on the availability of auxiliary data explaining the selectivity. In this sense, the availability of auxiliary variables explaining skewness in a data set is not very different when comparing survey response to secondary data sources.

Finally it must be noted that none of the algorithms in this section have been optimized. The design-based estimator could benefit from using the same covariates as used in the other methods. For the model based estimator no model selection has occurred potentially leading to a better model. Similarly, the CART and ANN methods have not been optimized with respect to number of branches and number of hidden neurons, respectively. The results in this section simply illustrate that methods that are conceptually very different, can lead to very similar results.

5. Real world applications

5.1 Social media data

Some experimental work has been conducted regarding social media messages. This relates to sociological research aimed at measuring peoples' opinions, sometimes

referred to as opinion mining. Since Twitter is used very actively in the Netherlands, Twitter messages of Dutch users were studied in detail (Daas et al., 2011; 2012).

Preprocessing of Twitter messages is a transformation of the list of messages into a list of people, with as a relevant survey variable, for example, their main topic of interest in a given week. Inference would then consist of predicting the topics of interest of the non-twittering population in the Netherlands. The missing link to do this is the lack of auxiliary information within the Twitter data set. Inference on just the available data really is inference about the Dutch Twitter population, which may be interesting, but in the setting of an NSI it is probably of no immediate relevance. A potential source of auxiliary data is the profile information provided by users.

Regardless of this, the first challenge is preprocessing: the extraction of relevant features from the short textual information in a Twitter message. Methods aimed at (pre)processing of textual data are often jointly referred to as text mining. This approach was used to classify and compare the topics discussed in Dutch Twitter messages with the various themes Statistics Netherlands is interested in. Automatic and manual classification results revealed that close to 55% of the Dutch Twitter messages collected could be allocated to themes of interest. The other 45% were allocated to a single 'remainder' group and not analyzed further. Classification of messages proved challenging. This was, amongst others, caused by the distorting effect of the large number of Twitter messages that are not relevant for the purpose of official statistics (Daas et al., 2012).

5.2 Mobile phone data

A set of mobile phone call records is available from a provider, containing phone ID, date, time and originating location of the calls made through that provider over a two week period (De Jonge et al., 2012). Preprocessing in this case requires the transformation of phone calls to a list of units of interest and some associated target variables. The units of interest are people, and the variables of interest can include their home and work location, the distance between the two, the number of trips or the total distance travelled. Phone call records can be used to derive likely home and work locations for some of the people/phones in the available data set.

Until now, in the data set used for experimentation, no identifying information was available about the phone owners. Assuming the information will be available in future, names and addresses can be linked with the population and employment registers, so deriving the subset of employed people, with some background characteristics as age, gender, income (from the tax office), etc. The mobile phone data would thus make available the home-work distances of a subset of those people. Predictive inference methods can then be used to estimate home-work distances for the other people.

The risk in doing this is that the people who are customers of a specific provider, and who make phone calls allowing determination of home and work location, are a selective group, differing in some way from the others. The key element is, as

discussed above, that the auxiliary data used in the estimators must explain such selectivity. If they do not, the resulting estimate may be biased.

The degree to which auxiliary data explain selectivity is difficult to assess, but can be determined in some occasions. For example, comparing results with outcomes from the mobility survey could be informative. In this survey, people report trips they make on a particular day. Notably, response levels in this survey are well below 50%, requiring auxiliary variables explaining participation in the survey sufficiently well. This too, can be problematic.

6. Conclusions

Investigation of tertiary data within NSIs is being conducted, with the ultimate aim of using such data sets to produce official statistics. A key element is the method of inference used to estimate population quantities. Since these data sets are not the result of a sample survey, they have no underlying sampling design. Hence, typical design based estimation methods are not applicable. Model-based estimation methods may provide a suitable alternative, but may be limited in their capabilities of predicting target variables from known background variables. Therefore widening the scope of predictive inference to include algorithmic, non-parametric methods is likely to be beneficial in some instances.

Boundaries between design-based and predictive inference have been fading for some time. Not only has there been the adoption of modelling approaches in survey settings (Pfeffermann, 2011), but more recently non-parametric methods have been applied successfully for inference using sample surveys too. Breidt et al. (2005), for example, use penalised splines for inference in complex surveys. This illustrates that paradigms are shifting. Techniques for inference used for official statistics will need to make this shift too, as the use and role of large scale sample surveys will diminish. This is instigated by the push to reduce data collection costs, and at the same time the rising availability of relatively inexpensive secondary and tertiary data.

The use of such alternative data sources in official statistics is in its infancy. Some important areas for research are the following:

- Determining relevant data sources and suitable preprocessing algorithms.
- Establishing ways to link these with statistical registers, or other methods to extend the data sets with auxiliary variables through data integration approaches.
- Determining and applying suitable methods of inference.
- Accurate error budgeting, including preprocessing, sampling and measurement errors.
- Independently validating the whole approach, possibly through one-off experiments or parallel surveys.

Apart from these methodological issues, it is imperative that official statistics that are produced in such novel ways comply with (inter)national standards and guidelines, and with privacy legislation.

While there seems to be an abundance of data in the world around us (Roos et al., 2009), turning a particular data source into a reliable source for official statistics is a rather elaborate process (Daas et al., 2011). Therefore careful consideration of potential gains and benefits must be made prior to starting such an undertaking. Nevertheless, exploiting the vast amounts of data out there in a smart and methodologically responsible way may open up new possibilities to produce fast, cheap, reliable and accurate statistics.

References

Andridge, R.R. and Little, R.J.A. (2011). Proxy Pattern-Mixture Analysis for Survey Nonresponse. *Journal of Official Statistics*, Vol. 27, No. 2, pp. 153-180.

Bakker, B. (2011). Micro integration. Statistical Methods paper 201108, Statistics Netherlands, the Hague/Heerlen.

Bethlehem, J.G. (1983). Ontwerpen of modelleren? Een discussie over het gebruik van modellen in de steekproeftheorie. CBS-rapport, Centraal Bureau voor de Statistiek, Voorburg.

Bethlehem, J.G., Cobben F. and Schouten B. (2011). *Handbook of nonresponse in household surveys*, Wiley.

Breidt, F., Claeskens, G. and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, Vol. 92, No. 4, pp. 831-846.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, Vol. 16, No. 3, pp. 199-231.

Daas, P.J.H., Roos, M., de Blois, C., Hoekstra, R., ten Bosch, O. and Ma, Y. (2011) New data sources for statistics: Experiences at Statistics Netherlands. Discussion paper 201109, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Daas, P.J.H., Roos, M., van de Ven, M. and Neroni, J. (2012) Twitter as a potential data source for statistics in the Netherlands. Paper presented at the 67th American Association for Public Opinion Research conference, Orlando, Florida, USA.

De Jonge, E., van Pelt M. and Roos, M. (2012). Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. Paper for the Federal Committee on Statistical Methodology research conference, Washington, USA.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification* (2nd edition). Wiley, New York.

Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository, available at: http://archive.ics.uci.edu/ml/.

Gelman, A. (2007). Struggles with survey weighting and regression modelling. *Statistical Science*, Vol. 22, Issue 2, pp. 153-164.

Hastie, T., Tibshirani R. and Friedman J. (2003). *The elements of statistical learning; data mining, inference, and prediction*, Second Ed., Springer.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, Vol. 47, No. 1, pp. 153-161.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* Vol. 47, pp. 663–685.

Kruskal, W., Mosteller, F. (1980) Representative Sampling, IV: the History of the Concept in Statistics, 1895-1939. *Int. Stat. Rev*., 48, 169-195.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. Journal of the American Statistical Association, 99, pp. 546-556.

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., and Ford, W. B. (1994). The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait. Technical Report 48, Marine Research Laboratories, Tasmania, Australia.

Nature (2008). *Big data*. Specials and supplements archive, http://www.nature.com/news/specials/bigdata/index.html, accessed 7 March 2012.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, Vol. 37, No. 2, pp. 115-136.

Pyle, D. (2009) *Data preparation for data mining*, Morgan Kaufmann, San Francisco.

Rao, J.N.K. (2003). *Small area estimation*. John Wiley & Sons, New-York.

Roos, M.R., Daas, P.J.H. and Puts, M. (2009) Innovative data collection: new data sources and opportunities (in Dutch). Discussion paper 09027, Statistics Netherlands, The Hague/Heerlen.

Sarndal, C.E., Swensson B. and Wretman J. (1992). *Model-assisted survey sampling*, Springer-Verlag, New-York.

Science (2011). *Dealing with data*. Special issue and online collection, http://www.sciencemag.org/site/special/data, accessed 7 March 2012.

Valliant, R., Dorfman A.H. and Royall R.M. (2000). *Finite Population Sampling and Inference – A Prediction Approach*. John Wiley & Sons, New-York.

Van den Brakel, J. (2012). Models in official statistics. Inaugural lecture, School of Business and Economics, Maastricht University.

Van den Brakel, J. and Bethlehem, J. (2008). Model-based estimation for Official Statistics. Technical Report 08002, Statistics Netherlands, Voorburg/Heerlen.

Wallgren A. and Wallgren B. (2007). *Register-based statistics: administrative data for statistical purposes*. Wiley.

Zhang, L.C. (2012a). Modeling of domain mortality rates, In ESSnet on Small Area Estimation, Report on Work Package 5, Case Studies. Eurostat, Luxembourg.

Zhang, L.C. (2012b). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.