Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination:

a simulation study based on the SBS sampling design at Statistics Belgium

Maria Caterina Bramati

FPS Economics -Belgium

NTTS 2009
Key issues on business surveys:

- harmonization of definitions (statistical, economic, fiscal...)
- harmonization of statistical methods
- centralization of data editing and processing
- IT support and standardization of statistical tools
- monitoring of the survey process
- data quality
Why integration?

Key issues on business surveys:

- harmonization of definitions (statistical, economic, fiscal...)
- harmonization of statistical methods
- centralization of data editing and processing
- IT support and standardization of statistical tools
- monitoring of the survey process
- data quality
Why integration?

Key issues on business surveys:

- harmonization of definitions (statistical, economic, fiscal...)
- harmonization of statistical methods
- centralization of data editing and processing
- IT support and standardization of statistical tools
- monitoring of the survey process
- data quality
Why integration?

Key issues on business surveys:

- harmonization of definitions (statistical, economic, fiscal...)
- harmonization of statistical methods
- centralization of data editing and processing
- IT support and standardization of statistical tools
- monitoring of the survey process
- data quality
Why integration?

Key issues on business surveys:

- harmonization of definitions (statistical, economic, fiscal...)
- harmonization of statistical methods
- centralization of data editing and processing
- IT support and standardization of statistical tools
- monitoring of the survey process
- data quality
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination:

Introduction

Why integration?

Key issues on business surveys:

- harmonization of definitions (statistical, economic, fiscal...)
- harmonization of statistical methods
- centralization of data editing and processing
- IT support and standardization of statistical tools
- monitoring of the survey process
- data quality
Introduction

Why integration?

Integration → 1 coherent variable-oriented structure driven at 4 levels:

- architecture
- definitions
- methods
- data flows
Pros and Cons

Advantages
-reduction of internal costs: several economies of scale in collection, data-processing, IT implementation, monitoring...
-synergies and cooperation between statisticians, sharing knowledge and expertise
-transparency of the procedures and improved documentation
-systematic approach to data quality
-monitoring and reduction of statistical burdens for businesses.

Drawback
-It requires a complete re-engineering of the statistical process: 25 surveys into 1!
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination:

Introduction

Structure of IBS

Main steps of statistical production:

- Universe, statistical unit and target population
- Sampling
- Collection
- Editing and data treatment
- Estimation
- Dissemination and quality reports
- ...
Main steps of statistical production:

- Universe, statistical unit and target population
- Sampling
- Collection
- Editing and data treatment
- Estimation
- Dissemination and quality reports
- ...
Structure of IBS

Main steps of statistical production:

- Universe, statistical unit and target population
- Sampling
- Collection
- Editing and data treatment
- Estimation
- Dissemination and quality reports
- ...
Sampling: 3 main problems

1. choice of the sampling design
2. sample size determination
3. sample allocation

under some constraints
- target variable(s) considered
- legal obligations and requirements (EUROSTAT, NBB, ...)
- available auxiliary information
Sampling: 3 main problems

1. choice of the sampling design
2. sample size determination
3. sample allocation

under some constraints
- target variable(s) considered
- legal obligations and requirements (EUROSTAT, NBB, ...)
- available auxiliary information
Sampling: 3 main problems

1. choice of the sampling design
2. sample size determination
3. sample allocation

under some constraints

- target variable(s) considered
- legal obligations and requirements (EUROSTAT, NBB, ...)
- available auxiliary information
Sampling: 3 main problems

1. choice of the sampling design
2. sample size determination
3. sample allocation

under some constraints

-target variable(s) considered
-legal obligations and requirements (EUROSTAT, NBB, ...)
-available auxiliary information
Sampling design: *stratified* sample

Main idea:
- population is divided into subgroups (or strata) in order to maximize the *intra*-group ‘*homogeneity*’ (according to a chosen target variable) and to minimize the *inter*-group ‘*homogeneity*’.
Sampling design: *stratified sample*

Main idea:
- population is divided into subgroups (or strata) in order to maximize the *intra*-group ‘*homogeneity*’ (according to a chosen target variable) and to minimize the *inter*-group ‘*homogeneity*’.

It requires
- mutually exclusive strata: 1 unit can belong to 1 stratum only
- collectively exhaustive strata: no population unit excluded
Quality concerns?

Ideally, the choice of 1) — 3) should be linked to quality issues of the final *statistical product*, balancing costs and benefits.
Quality concerns?

Ideally, the choice of 1) — 3) should be linked to quality issues of the final *statistical product*, balancing costs and benefits.

⇒ Target *statistical precision* is the constraint under which choices are made.
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination:
Sampling Methods for the IBS

Solution proposed: HL algorithm

The HL algorithm with Neyman allocation represents an optimal solution for the three problems.

\[
n_{\hat{t}_{ystrat}} = N_L + \sum_{h=1}^{L-1} \frac{W_h^2 s_{yh}^2}{a_h} \left( \frac{c Y / N}{(c Y / N)^2 + \sum_{h=1}^{L-1} \frac{W_h s_{yh}^2}{N}} \right) + \sum_{h=1}^{L-1} \frac{W_h s_{yh}^2}{N}
\]

(1)

\[
a_h = \frac{n_h}{N_h} = \frac{W_h s_{yh}}{\sum_{k=1}^{L-1} W_k s_{yk}}
\]

(2)
The idea of HL algorithm is to find the optimal strata boundaries $b_1, \ldots, b_{L-1}$ which minimize the size $n_{\hat{y}_{strat}}$ subject to a required precision $c$, with some appropriate sampling allocation (Neyman, proportional...).
The idea of HL algorithm is to find the optimal strata boundaries $b_1, \ldots, b_{L-1}$ which minimize the size $n_{\hat{t}_{y,\text{strat}}}$ subject to a required precision $c$, with some appropriate sampling allocation (Neyman, proportional...).

However
Solution proposed: HL algorithm

The idea of HL algorithm is to find the optimal strata boundaries \( b_1, \ldots, b_{L-1} \) which minimize the size \( n_{tystrat} \) subject to a required precision \( c \), with some appropriate sampling allocation (Neyman, proportional...).

However

1. \( s^2_{yh} \) is unknown \( \Rightarrow \) use of auxiliary information \( X \) for \( Y \)
2. number \( L \) of strata is selected by the user
Solution proposed: HL algorithm

The idea of HL algorithm is to find the optimal strata boundaries \( b_1, \ldots, b_{L-1} \) which minimize the size \( n^\text{tystrat} \) subject to a required precision \( c \), with some appropriate sampling allocation (Neyman, proportional...).

However

1. \( s^2_{yh} \) is unknown \( \implies \) use of auxiliary information \( X \) for \( Y \)
2. number \( L \) of strata is selected by the user

BUT

auxiliary information \( X \neq Y \) target variable.
Solution proposed: HL algorithm

⇒ modified HL algorithm:
the discrepancy existing between $Y$ and $X$ is estimated!
Solution proposed: HL algorithm

→ modified HL algorithm: the discrepancy existing between $Y$ and $X$ is estimated!

**Advantages [methods]**

- **statistical quality** approach (i.e. sample size chosen on quantitative grounds)
- optimal stratification and allocation
- discrepancies between $Y$ and $X$ are modelled
Solution proposed: HL algorithm

→ modified HL algorithm: the discrepancy existing between $Y$ and $X$ is estimated!

Advantages [methods]

- **statistical quality** approach (i.e. sample size chosen on quantitative grounds)
- **optimal** stratification and allocation
- discrepancies between $Y$ and $X$ are modelled
Solution proposed: HL algorithm

→ modified HL algorithm: the discrepancy existing between $Y$ and $X$ is estimated!

Advantages [methods]

- **statistical quality** approach (i.e. sample size chosen on quantitative grounds)
- **optimal** stratification and allocation
- discrepancies between $Y$ and $X$ are modelled
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination: Sampling Methods for the IBS

Solution proposed: **modified** HL algorithm

**Advantages** [technical]

- general procedure (of easy application to many surveys)
- fully documented (web resources, scientific literature)
- (fast) algorithm available in SAS
- possible further improvement (robustness, multiple survey variables approach...)

**Drawbacks**

- auxiliary variables needed
- optimality of the design achieved w.r.t. 1 single survey variable.
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination: Sampling Methods for the IBS

Solution proposed: **modified** HL algorithm

**Advantages [technical]**

- **general** procedure (of easy application to many surveys)
- fully documented (web resources, scientific literature)
- (fast) algorithm available in SAS
- possible further improvement (robustness, multiple survey variables approach...)

**Drawbacks**

- auxiliary variables needed
- optimality of the design achieved w.r.t. 1 single survey variable.
Solution proposed: modified HL algorithm

Advantages [technical]

- general procedure (of easy application to many surveys)
- fully documented (web resources, scientific literature)
- (fast) algorithm available in SAS
- possible further improvement (robustness, multiple survey variables approach...)

Drawbacks

- auxiliary variables needed
- optimality of the design achieved w.r.t. 1 single survey variable.
Solution proposed: modified HL algorithm

Advantages [technical]

- general procedure (of easy application to many surveys)
- fully documented (web resources, scientific literature)
- (fast) algorithm available in SAS
- possible further improvement (robustness, multiple survey variables approach...)

Drawbacks

- auxiliary variables needed
- optimality of the design achieved w.r.t. 1 single survey variable.
Solution proposed: modified HL algorithm

Advantages [technical]

- general procedure (of easy application to many surveys)
- fully documented (web resources, scientific literature)
- (fast) algorithm available in SAS
- possible further improvement (robustness, multiple survey variables approach...)

Drawbacks

- auxiliary variables needed
- optimality of the design achieved w.r.t. 1 single survey variable.
Solution proposed: modified HL algorithm

Advantages [technical]

- general procedure (of easy application to many surveys)
- fully documented (web resources, scientific literature)
- (fast) algorithm available in SAS
- possible further improvement (robustness, multiple survey variables approach...)

Drawbacks

- auxiliary variables needed
- optimality of the design achieved w.r.t. 1 single survey variable.
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination:
Sampling Methods for the IBS

Solution proposed: *modified* HL algorithm

**Advantages** [technical]

- **general** procedure (of easy application to many surveys)
- **fully documented** (web resources, scientific literature)
- (fast) algorithm available in SAS
- possible further improvement (robustness, multiple survey variables approach...)

**Drawbacks**

- auxiliary variables needed
- optimality of the design achieved w.r.t. 1 single survey variable.
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination: Simulation study
Ad-hoc stratification vs HL stratification
Simulation study

Description of the experiment:
Build a stratified sample for surveying value added (=y), by
Ad-hoc stratification vs HL stratification

Simulation study

Description of the experiment:
Build a stratified sample for surveying value added ($y$), by

- $i =$ NACE 4-digits class
- $j =$ economic-size class
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination: Simulation study

Ad-hoc stratification vs HL stratification Simulation study

Description of the experiment:
Build a stratified sample for surveying value added ($=y$), by

- $i =$ NACE 4-digits class
- $j =$ economic-size class

Population is generated from

$$\log y_{ij}^{\text{sim}} = \hat{\beta}_1 \log x_{ij}^{(1)} + \hat{\beta}_2 x_{ij}^{(2)} + \varepsilon_{ij}$$
Ad-hoc stratification vs HL stratification
Simulation study

Description of the experiment:
Build a stratified sample for surveying value added \((= y)\), by

- \(i = \text{NACE 4-digits class}\)
- \(j = \text{economic-size class}\)

Population is generated from

\[
\log y_{ij}^{\text{sim}} = \hat{\beta}_1 \log x_{ij}^{(1)} + \hat{\beta}_2 x_{ij}^{(2)} + \varepsilon_{ij}
\]

\(x^{(1)} = \text{turnover (VAT register)}\)
\(x^{(2)} = \text{number of employees (ONSS register)}\)
Ad-hoc stratification vs HL stratification
Simulation study

Description of the experiment:
Build a stratified sample for surveying value added ($=y$), by

- $i = \text{NACE 4-digits class}$
- $j = \text{economic-size class}$

Population is generated from

$$\log y_{ij}^{\text{sim}} = \hat{\beta}_1 \log x_{ij}^{(1)} + \hat{\beta}_2 x_{ij}^{(2)} + \varepsilon_{ij}$$

$x^{(1)} = \text{turnover (VAT register)}$
$x^{(2)} = \text{number of employees (ONSS register)}$
$\hat{\beta}_1, \hat{\beta}_2$ estimated using the SBS sampling frame
6 strata bounds on variable $x^{(1)}$ for each of the 4-digits NACE class ($= 1$ take-all + 5 take-some)
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination:
Simulation study

Ad-hoc stratification vs HL stratification
Simulation study

6 strata bounds on variable $x^{(1)}$ for each of the 4-digits NACE class
($= 1$ take-all + 5 take-some)

- ad-hoc size criterion used by the SBS
  vs
- generalized HL method at 1% precision using $x^{(1)}$ only as auxiliary
  information.
6 strata bounds on variable $x^{(1)}$ for each of the 4-digits NACE class
($= 1$ take-all + 5 take-some)

-ad-hoc size criterion used by the SBS vs
- generalized HL method at 1% precision using $x^{(1)}$ only as auxiliary information.

Stratification performance is evaluated on MSE basis.
Generalized Hidiroglou-Lavallée vs ad-hoc strata bounds determination: Simulation study

Ad-hoc stratification vs HL stratification

<table>
<thead>
<tr>
<th>Sector of Activities</th>
<th>Sample reduction</th>
<th>Gain in precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>−39%</td>
<td>+12%</td>
</tr>
<tr>
<td>Construction</td>
<td>−30%</td>
<td>+50%</td>
</tr>
<tr>
<td>Trade</td>
<td>−29%</td>
<td>+20%</td>
</tr>
<tr>
<td>Services</td>
<td>+15%</td>
<td>+61%</td>
</tr>
</tbody>
</table>

Table: Summary of results comparing modified HL method versus SBS ad-hoc stratification (based on average MSE).

Target precision: 1%
Sampling methods: agenda

Further improvements:

- tackle the problem of outliers in the auxiliary variable: robust issues
- stratification based on more than 1 target variable: multiple survey optimal stratification
- rotating panel for SMEs and statistical holidays
Sampling methods: agenda

Further improvements:

- tackle the problem of outliers in the auxiliary variable: robust issues
- stratification based on more than 1 target variable: multiple survey optimal stratification
- rotating panel for SMEs and statistical holidays
Further improvements:

- tackle the problem of outliers in the auxiliary variable: robust issues
- stratification based on more than 1 target variable: multiple survey optimal stratification
- rotating panel for SMEs and statistical holidays
Further improvements:

- tackle the problem of outliers in the auxiliary variable: robust issues
- stratification based on more than 1 target variable: multiple survey optimal stratification
- rotating panel for SMEs and statistical holidays