Selective editing of business investments by using administrative data as auxiliary information

Marco Di Zio (Istat)
Ugo Guarnera (Istat)
Massimiliano Iommi (Istat)
Antonio Regano (Istat)

Budapest, 14-16/09/2015
Outline

1. Motivations
2. Auxiliary information from administrative data
3. Selective editing using SeleMix
4. Application of SeleMix to investment data
5. Validation
6. Conclusions and future research
Motivations (1/5)

- Investments (Gross Fixed Capital Formation) is of great interest in National Accounts because:
  - They are an important component of final demand (and then of GDP)
  - They are an input to estimate capital stock and consumption of fixed capital (that are needed to produce productivity and profitability indicators)
- Traditionally National Accounts department in Istat makes some editing to SBS data (it is not a simple “user” of SBS data)
- However the selective editing procedure we describe is not “National Accounts specific”
Motivations (2/5)

• Administrative data are part of the process of SBS in Istat
  ✓ Istat currently has access to a rich set of administrative data regarding
    variables reported in profit and loss accounts and (only for corporations
    and limited companies) in balance sheets and to data from tax
    declarations.

• Investment data for the production of SBS for Italy are collected via a
  survey because no administrative data is available.

• Investment data are reported only in the Notes to Financial Statement
  ✓ Notes comprising a summary of significant accounting policies and details
    of the values reported in the profit and loss account and in balance sheet
Motivations (3/5)

• Istat has access to the explanatory notes of corporations and limited companies in two forms

• In the form of non-standardized text files (one for each company)
  ✓ A valuable source to recover the “true” value of investment for a limited number of firms

• In the form of an experimental dataset reporting the value of investment that is obtained using a software for automatic optical recognition from the non-standardized text files.
  ✓ It cannot be used to produce SBS data or to automatically correct the data because of the errors due to the automatic optical recognition
  ✓ Valuable source for selective editing
Motivations (4/5)

• Recovering investment data from the explanatory notes is a time consuming process.

• The goal of selective editing framework is just to minimize the number of checks by focusing on the units where editing has the highest expected benefit.

• Typically, selective editing is useful if at the review stage it is possible to recover with a high reliability the ‘true’ value

 ✓ In the case of investments, the revision of the values of the units is done by consulting the Explanatory Notes.
Motivations (5/5)

- Di Zio et al. (2014):
  - SeleMix
  - Investment of the same firm in the previous year as covariate
  - Notes as the source to recover the “true” value of investment.

- The results were overall satisfactory, but they also showed that using historical values as covariate for investment is far from ideal:
  - The hit rate (i.e., the percentage of erroneous observation in the selected units) was not very high (44.5%).
  - it is difficult to select an observation with an erroneous investment that it is not atypical with respect to the historical value (as can be the case when both the current and the historical variables are affected by the same error mechanism).
Auxiliary information from administrative data (1/4)

- A derived variable that can be calculated from the information in Financial Statement (FS hereafter) currently available to Istat (DELTA_STOCK variable hereafter)
- Expenditures for amortizable goods reported in Value Added Tax declarations (VAT variable hereafter)
- Total investment from the experimental database on the explanatory note (EXNOTE variable hereafter)
Auxiliary information from administrative data (2/4)

- In companies’ financial statements, acquisition of fixed assets (i.e. investment) is one of the components that explain the difference between the value of net asset at the beginning of the accounting period and the value at the end of the accounting period:

\[ \text{NETSTOCK}_{t,\text{end}} = \text{NETSTOCK}_{t,\text{beg}} + \text{INVE}_t + \text{REVAL}_t - \text{SOLD}_t - \text{AMOR}_t - \text{WOFF}_t + \text{MA}_t \]

- \( \text{NETSTOCK}_{t,\text{end}} \) are net assets at the end and of year \( t \),
- \( \text{NETSTOCK}_{t,\text{beg}} \) are net assets at the end beginning of year \( t \),
- \( \text{INVE}_t \) is the acquisition of assets (investment) in year \( t \),
- \( \text{REVAL}_t \) is the revaluation of existing assets in the year \( t \),
- \( \text{SOLD}_t \) is the net book value (i.e., net of cumulated depreciation) of existing assets sold in the year \( t \),
- \( \text{AMOR}_t \) is depreciation of existing assets in the year \( t \),
- \( \text{WOFF}_t \) is write-off of existing assets in the year \( t \)
- \( \text{MA}_t \) is the net effect of mergers and acquisitions in year \( t \).
Selective editing of business investments by using administrative data as auxiliary information, Di Zio et al, Budapest, 14-16/09/2015

Auxiliary information from administrative data (3/4)

From FS database we can compute:

\[ \text{NETSTOCK}_{t,\text{end}} - \text{NETSTOCK}_{t,\text{beg}} + \text{AMOR}_t + \text{WOFF}_t = \text{DELTA\_STOCK}_t \]

\[ \text{INVE}_t + \text{REVAL}_t - \text{SOLD}_t + \text{MA}_t \]
Selective editing using SeleMix (1/3)

- The selective editing method is based on explicitly modelling both true (error-free) data and error mechanism (see Di Zio and Guarnera, 2013, for details).

- The model assumptions can be summarized as follows
  - True data are thought of as n realizations from a random p-vector $Y$ that, conditional on a set of q covariates $X$ is normally distributed with mean vector $BX$ and covariance matrix $\Sigma$
  - The intermittent nature of the error is modelled through a Bernoullian r.v
  - Conditional on the presence of error, we assume a Gaussian additive error with zero mean and covariance matrix proportional to $\Sigma$

- The distribution of true data conditional on observed data is a mixture of a mass density corresponding to absence of error and a Gaussian distribution corresponding to presence of error
  - The model parameters can be estimated by maximizing the likelihood function based on the observed data
Selective editing using SeleMix (2/3)

• The selective editing strategy consists in using the estimated distribution of true data conditional on observed data to build up a score function.

• Score function = difference between observed and anticipated value
  ✓ The anticipated value is obtained by means of a weighted average of the observed value and a synthetic value.
    o The weights are given by the probability of being in error.
    o The synthetic value is in turn the weighted average of the observed value and a robust estimate of the regressed value.

• Once all the observations have been ordered according to this score function, it is possible to estimate the residual error remaining in data after the correction of the first k units (k=1,..,n).

• The number of most critical units to be edited can be chosen so that the estimate of the residual error is below a prefixed threshold
  • Selemix allows to explicitly relate the efforts in editing activities to the accuracy of the target estimates.
Selective editing using SeleMix (3/3)

\( Y^* \) true data

\( Y \) observed data

\( X \) covariates (no error)

\( B \) regression coefficients

\( U \) residuals

\( I \) Bernoullian variable:

\[
\pi = P\{I = 1\} = 1 - P\{I = 0\}
\]

**True data model:**

\[
Y^* = XB + U, \quad U \sim N(0, \Sigma)
\]

**Error model:**

\[
P\{Y \neq Y^*\} = P\{I = 1\} = \pi,
\]

\[
Y \mid Y^*, I = Y^* + I \varepsilon
\]

\[
\varepsilon \sim N(0, \Sigma_\varepsilon) \quad \Sigma_\varepsilon = (\alpha - 1)\Sigma, \quad \alpha > 1.
\]

**Distribution of observed data:**

\[
f_Y(y) = (1 - \pi)N(y; B' x, \Sigma) + \pi N(y; B' x, \alpha \Sigma)
\]
Application of SeleMix to investment data (1/5)

• Data source: Istat Annual Survey on Economic and financial accounts of large enterprises, year 2011
  ✓ Actually a census, covering all enterprises operating in Italy with at least 100 employees
  ✓ Industrial and services sectors excluding financial services
  ✓ EC Structural Business Statistics.

• Units: responding enterprises (5770 observations)

• Target variable: total gross investment

• Covariates: DELTA_STOCK, VAT, EXNOTE
  ✓ Covariates are considered not affected by errors
Application of SeleMix to investment data (2/5)

- Estimation domains:
  - Model estimated separately for each NACE Rev2 section
  - Impact of errors evaluated at the level of 64 industries
    - corresponding to the classification of economic activity A*64 that is used to disseminate National Accounts data
  - Threshold used for the selection of critical units is chosen such that the estimated relative residual error for each industry is 4%
    - threshold chosen taking into account the reasonable number of units that can be checked with the available resources
  - The model is estimated on the subset of data with positive values of the analyzed variables
    - predictions are computed also for units whose observed value is zero
  - A weight for taking into account the non-response is computed and used in the score function
Pre-processing is needed to distinguish between missing values and “genuine zeros” in the target variable

- In fact, in the survey missing values are registered with a ‘zero’
- By considering as “zero investments” all cases where no positive values are reported, could result in an underestimation of the aggregates
- The underestimation is given by counting as zeros the missing values having an investment

- We decided to consider missing values all cases where the target variable is zero and at least one of the three variables VAT, DELTA_STOCK and EXNOTE are greater than zero
Application of SeleMix to investment data (4/5)

- Treatment of different missing pattern for the covariates
  - In fact for different units different sets of covariates may be available.
  - We estimate different models for different patterns, where, for each pattern, the units included in the estimation process are all those where (at least) the covariates corresponding to the current pattern are available.
    - For instance, if the current pattern corresponds to the availability of DELTA_STOCK and EXNOTE, the data used to estimate the parameters are all those having the first DELTA_STOCK and EXNOTE observed, regardless of whether VAT is observed or not
    - The estimated parameters are used to predict the errors only in the data with DELTA_STOCK and EXNOTE observed.
Application of SeleMix to investment data (5/5)

<table>
<thead>
<tr>
<th>N. Obs.</th>
<th>Selected</th>
<th>% Selected</th>
<th>Edited</th>
<th>Hit-Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5770</td>
<td>106</td>
<td>1.8%</td>
<td>77</td>
<td>72.6%</td>
</tr>
</tbody>
</table>

- Hit-rate quite high and the number of selected units very low

<table>
<thead>
<tr>
<th>SBS Original Value (A)</th>
<th>Post-editing Value (B)</th>
<th>(B - A)</th>
<th>(B - A) / A * 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>44,473 mlnE</td>
<td>45,576 mlnE</td>
<td>1,103 mlnE</td>
<td>-2.5%</td>
</tr>
</tbody>
</table>

- Notwithstanding the low number of selected units, not a negligible impact on the aggregate estimated value: efficient procedure
- Very high impact at the industry level (ranging from -47% to 118%)
Validation (1/3)

• The assessment of the quality of the procedure cannot be based only on its efficiency
• It is also necessary to quantify the impact of the residual errors on the estimates.
• The ideal validation of the procedure would consist in revising all the not selected observations of the sample by looking at their reported values in the notes
  ✓ We could find the error left in each not selected unit (residual error)
  ✓ We could compute the impact of residual errors on the released estimate
• The manual revision of all not selected observations is not feasible in practice unless a big investment in resources is planned.
• Alternative strategy: Manual revision of all the not selected observations in some industries
  ✓ The analysis of the results in each industry is quite informative, because the threshold used (set equal to 4%) is referred to each industry
Validation (2/3)

<table>
<thead>
<tr>
<th>Industry</th>
<th>N. Obs</th>
<th>N. Selected obs.</th>
<th>N. Edited Obs.</th>
<th>RE - raw data</th>
<th>RE - post editing</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 - Water transport</td>
<td>30</td>
<td>3</td>
<td>3</td>
<td>-35.3%</td>
<td>2.0%</td>
</tr>
<tr>
<td>38 - Motion picture</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>-0.2%</td>
<td>-0.1%</td>
</tr>
<tr>
<td>39 - Telecommunications</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>26.05%</td>
<td>26.1%</td>
</tr>
</tbody>
</table>

- Manually revised three industries
- Industries 32 and 38 satisfactory results
  - Post editing residual percent error (RE) lower to the expected one (4%)
  - In industry 32 strong improvement obtained selecting few units
  - In industry 38 correctly no observation has been selected as influential
- Industry 39 not satisfactory:
  - The selective procedure does not select any observation
  - the error is 26.1%.
Validation (3/3)

• What did happen in industry 39?
• The error is due almost entirely to one company that was not selected by the editing procedure:
• The company did not report correctly in SBS questionnaire the value of a reclassification from assets in progress and advances to the corresponding assets categories.
• Why the selective editing procedure did not select this unit?
  ✓ EXNOTE is not observable for this firm
  ✓ DELTA_STOCK is very different from the true value of investment because of the effect an important Merger and Acquisition activity.
  ✓ Unfortunately, the value of DELTA_STOCK is not very far from the (wrong) value reported in the SBS questionnaire, hence the model cannot see any discrepancy between target variable and covariate
  ✓ There is a discrepancy between the target variable and the other covariate observed for this firm (i.e. VAT). The model gave more weight to the similarity of DELTA_STOCK than to the discrepancy of VAT.
Summing up

- The selective editing procedure proved to be quite efficient: strong improvements in the results obtained selecting few units and the hit rate is quite high.
- The selective editing procedure is considered by the survey managers feasible in terms of costs.
- The use of covariates specific for investment that are available from administrative data improves the efficiency of the procedure with respect to the use of historical values.
- The analysis performed to validate the method is encouraging and should be extended to other estimation domains.
- A result of the validation analysis is that the explanatory power of DELTA_STOCK is not satisfactory for firms that have been involved in mergers and acquisitions.
Future developments

- Using as additional explanatory variable also information on mergers and acquisitions that are available from business register
- Experimenting with an alternative validation procedure, based on the use of samples of not selected observations.
  - PPS sampling proportional to the scores estimated with SeleMix
- Revising all the observations belonging to the domains where the residual error estimated with the previous scheme is considerably far from the prefixed threshold
  - ✔ Removing errors from the historical values
  - ✔ Understanding whether there are other important error mechanisms affecting the quality of the procedure.
Thank you