Mind the Gaps: Explaining missing and poor quality milex data

Sam Perlo-Freeman (SIPRI) & Neil Ferguson (???)
19th International Conference on Economics & Security, Grenoble 2015
Problems with milex data

• The SIPRI milex dataset (like all others) has a lot of gaps in it.
• Some countries keep military expenditure figures a secret. Others may make it public in principle, but may be very hard to find.
• Even where data exists, it may be available only in highly aggregated form, or unreliable due to off-budget spending or lack of actual expenditure figures.
Transparency in military spending

- Military budgeting and expenditure transparency often weak, more than other budget areas
- National security excuse, exemption of military from normal processes and oversight
- Transparency and accountability includes both provision of information and openness of processes
Significance of transparency

• For citizens to know how their money is being spent
• Lack of transparency creates opportunities for corruption
• *Externally*, transparency in military affairs can contribute to confidence-building measures
• For econometricians, systematic patterns in data gaps may generate sample-selection bias
Determinants of data availability & quality

- This study estimates the determinants of milex data availability and quality from 1995-2013 using duration-based and multiple logit models.

- We consider two main types of determinant:
  - *Openness*: willingness to provide data
  - *Capacity* of institutions to produce and distribute reliable data
Potential determinants

- Democracies may be more likely to produce detailed budget info, as accountable to populations.
- Autocracies may be more paranoid about information on military affairs.
- State fragility – states that have collapsed due to conflict, or that have weak institutions, may lack capacity to produce and distribute data.
- Income another potential capacity factor.
Coding data quality

- Coded available military expenditure data in SIPRI database from 1988-2014 to 4 levels:
  - 0 = No data
  - 1 = Secondary sources only
  - 2 = Primary source, ‘low’ quality
  - 3 = Primary source, medium/high quality

- For category 3, required breakdown by major category (personnel, O&M, equipment etc.), actual expenditure data, no major known or strongly suspected off-budget spending

- Some uncertainty in coding between 2 and 3
Data quality 1988-2014

Military expenditure data availability, world, 1988-2014

- No data
- Secondary data
- Primary, low qual
- Primary, med/high qual
Data quality 1988-2014 Africa
Data quality 1988-2014 Asia

Milex data availability, Asia & Oceania 1988-2014

- No data
- Secondary data
- Primary, low qual
- Primary, med/high qual
Data quality 1988-2014 C & E Europe
Data quality 1988-2014 Latin America

Milex data availability, Latin America & the Caribbean 1988-2014

- No data
- Secondary data
- Primary, low qual
- Primary, med/high qual
Data quality 1988-2014 Middle East

Milex data availability, Middle East 1988-2014
Independent variables

• Democracy: Polity IV dataset of democracy and autocracy, scale of -10 to +10.
  – Adjusted version of Polity2 score, interpolating ‘transition’ cases and giving value of 0 to ‘interrupted’ and ‘collapsed’ polities.

• Fragility: World Bank fragility index from 1995, scale of 0 to 25: conflict, weak institutions, presence of foreign troops, etc.

• World bank income group: low, lower-middle, upper-middle, high.
Independent variables

• Democracy: Polity IV dataset of democracy and autocracy, scale of -10 to +10.
  – Adjusted version of Polity2 score, interpolating ‘transition’ cases and giving value of 0 to ‘interrupted’ and ‘collapsed’ polities.

• Fragility: World Bank fragility index from 1995, scale of 0 to 25: conflict, weak institutions, presence of foreign troops, etc.

• World bank income group: low, lower-middle, upper-middle, high.

• Non-OECD oil producer dummy; regional dummies
Average value of indep. vars
Model

• Used Complementary log-log models to provide best discrete approximation to Cox proportional hazard model

• Duration effect obtained by including dummies for each year, and a linear time variable.

• Three binary models: “some data” (1-3) vs “no data” (0); “primary data” (2-3) vs no primary data (0-1); and “good data” (3) vs. less good data (0-2).

• Also multinomial logit model with all 4 values. Inverted quality scale to make best data the base case, and the others different levels of ‘failure’
## Results of C-log-log models

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coeff.</th>
<th>S.E.</th>
<th>Coeff.</th>
<th>S.E.</th>
<th>Coeff.</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No data</td>
<td></td>
<td></td>
<td>No primary</td>
<td></td>
<td>Not good</td>
<td></td>
</tr>
<tr>
<td>Polity2</td>
<td>-.124***</td>
<td>.030</td>
<td>-.083***</td>
<td>.021</td>
<td>-.095***</td>
<td>.017</td>
</tr>
<tr>
<td>Fragility</td>
<td>.110***</td>
<td>.037</td>
<td>.063**</td>
<td>.026</td>
<td>.075**</td>
<td>.029</td>
</tr>
<tr>
<td>Income group</td>
<td>-.215</td>
<td>.283</td>
<td>-.718***</td>
<td>.248</td>
<td>-.352**</td>
<td>.168</td>
</tr>
<tr>
<td>Non-OECD oil</td>
<td>.781</td>
<td>.83</td>
<td>2.56***</td>
<td>.704</td>
<td>1.26***</td>
<td>.416</td>
</tr>
<tr>
<td>% ‘failures’</td>
<td>8.7%</td>
<td></td>
<td>22.7%</td>
<td></td>
<td>59.8%</td>
<td></td>
</tr>
<tr>
<td>Pseudo log-like.</td>
<td>-697</td>
<td></td>
<td>-1134</td>
<td></td>
<td>-1196</td>
<td></td>
</tr>
</tbody>
</table>

- Results not significantly different for logit and multiple logit
Implication of coefficients

• ‘Typical’ case: Polity2 = 3, Fragility = 9, income group = lower-middle

<table>
<thead>
<tr>
<th></th>
<th>No data</th>
<th>No primary</th>
<th>Not best quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical case</td>
<td>8.7%</td>
<td>22.7%</td>
<td>59.8%</td>
</tr>
<tr>
<td>Least democratic</td>
<td>36.7%</td>
<td>53.1%</td>
<td>95.6%</td>
</tr>
<tr>
<td>Most democratic</td>
<td>3.8%</td>
<td>13.4%</td>
<td>37.4%</td>
</tr>
<tr>
<td>Most fragile</td>
<td>41.2%</td>
<td>50.6%</td>
<td>95.1%</td>
</tr>
<tr>
<td>Least fragile</td>
<td>3.3%</td>
<td>13.6%</td>
<td>37.1%</td>
</tr>
<tr>
<td>Low income</td>
<td>10.7%</td>
<td>41.0%</td>
<td>84.2%</td>
</tr>
<tr>
<td>High income</td>
<td>5.8%</td>
<td>5.9%</td>
<td>47.3%</td>
</tr>
</tbody>
</table>
Conclusions

• Strong and robust impact of democracy on data availability and quality
• State fragility likewise has a strong and generally robust impact, indicating capacity as well as willingness may affect data availability
• Income appears not to have a significant affect on minimal data availability, but does have an effect on data quality.