The case of bounds in noisy protection methods: Selected risk and utility perspectives from official population statistics

2023 UNECE Expert Meeting on SDC, 26 – 28 September 2023
Risk assessment: Privacy, confidentiality, and disclosure vs utility

Fabian BACH
European Commission – Eurostat
Unit F2 – Population and migration
Outline

1. Intro: Noisy methods and bounds in a nutshell
2. Specific utility flaws of *unbounded* noise
3. Specific additional disclosure risks of *bounded* noise
4. Conclusions
Intro: noisy methods and bounds in a nutshell

- SDC $\leftrightarrow$ protect individuals

<table>
<thead>
<tr>
<th>SEX \ POB*</th>
<th>Total</th>
<th>Country</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>42</td>
<td>35</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>22</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>Female</td>
<td>20</td>
<td>18</td>
<td>2</td>
</tr>
</tbody>
</table>

* Place of birth (POB)
Intro: noisy methods and bounds in a nutshell

- SDC $\leftrightarrow$ protect individuals
- old-school suppression often inefficient and inconsistent

<table>
<thead>
<tr>
<th>SEX \ POB</th>
<th>Total</th>
<th>Country</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>42</td>
<td>35</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>22</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Female</td>
<td>20</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>
Intro: noisy methods and bounds in a nutshell

- SDC $\leftrightarrow$ protect individuals
- old-school suppression often inefficient and inconsistent
- Noise in action: Is this better?

<table>
<thead>
<tr>
<th>SEX \ POB</th>
<th>Total</th>
<th>Country</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>42</td>
<td>37</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>23</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>
# Intro: noisy methods and bounds in a nutshell

- SDC $\leftrightarrow$ protect **individuals**
- Old-school suppression often inefficient and inconsistent
- Noise in action: **Is this better?**

... a closer look at a **single statistic** ...

<table>
<thead>
<tr>
<th>SEX \ POB</th>
<th>Total</th>
<th>Country</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>42</td>
<td>37</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>23</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>
Intro: noisy methods and bounds in a nutshell

... a closer look at single statistic level: intrinsic uncertainty

![Graph showing intrinsic uncertainty with a range of ±2]

\[
\text{intrinsic uncertainty: } \pm 2
\]
Intro: noisy methods and bounds in a nutshell

... a closer look at single statistic level: intrinsic uncertainty vs. noise

\[ \approx \]

probability

intrisic uncertainty: ± 2
protective noise added: ± 1
Intro: noisy methods and bounds in a nutshell

... a closer look at single statistic level: intrinsic uncertainty vs. noise

- intrinsic uncertainty: ± 2
- protective noise added: ± 1
- total uncertainty: ± 2.2
Intro: noisy methods and bounds in a nutshell

... a closer look at single statistic level: intrinsic uncertainty vs. noise

![Diagram showing probability distribution with intrinsic uncertainty, protective noise added, and total uncertainty.]

```
intrinsic uncertainty: ± 2
protective noise added: ± 2
total uncertainty: ± 2.8
```

“Damage” of noise protection can be gauged against intrinsic uncertainty.
Intro: noisy methods and bounds in a nutshell

... a closer look at single statistic level: **noise distributions**

noise variance $V$ (often) free parameter

protective noise added: $\pm 2$
Intro: noisy methods and bounds in a nutshell

- **Noise distributions**: how long is the tail?
Intro: noisy methods and bounds in a nutshell

• **Noise distributions**: how long is the tail?

![Diagram showing noise distributions and bounds](image)

- CKM bound parameter $E = 3$
- e.g. relaxed $(\epsilon,\delta)$-DP or cell key method (CKM)
- e.g. strict $\epsilon$-DP

... but **strict $\epsilon$-DP**
Intro: noisy methods and bounds in a nutshell

- **Noise distributions**: how long is the tail?

- Specific risk/utility issues related to the bound...

![Diagram showing noise distributions with CKM bound parameter $E = 3$ and relaxed $(\varepsilon, \delta)$-DP or cell key method (CKM) compared to strict $\varepsilon$-DP. The diagram illustrates the noise probability distribution and the deviation $\delta$ between bounds.]
Utility flaws of *unbounded* noise

- 2021 EU census: ca. 110 000 Local Administrative Units (~ municipalities), of which
  - 43 395 with <500 people
  - 8 502 with <100 people
  - 866 with <20 people
- Could we accept here e.g. $Pr(|\text{noise}|>100) = 0.1\%$ or more?
  - Yes
  - No
Utility flaws of *unbounded* noise

- 2021 EU census: ca. 110,000 Local Administrative Units (~ municipalities), of which
  - 43,395 with <500 people
  - 8,502 with <100 people
  - 866 with <20 people
- Could we accept here e.g. $\Pr(|\text{noise}|>100) = 0.1\%$ or more?
  - ❑ Yes
  - ❑ No
Utility flaws of *unbounded* noise: counts

- E.g. 2020 U.S. census test setup with moderate tabular $\varepsilon = 0.1$
- expectation for individual LAU counts to obtain noise of relative size $\pm 20$, $\pm 50$ and $\pm 100$
- analytical estimation (bins) and numerical simulation (lines)
Utility flaws of *unbounded* noise: counts

- E.g. 2020 U.S. census test setup with restrictive tabular $\varepsilon = 0.025$
- expectation for individual LAU counts to obtain noise of relative size $\pm 20$, $\pm 50$ and $\pm 100$
- analytical estimation (bins) and numerical simulation (lines)
Utility flaws of *unbounded* noise: counts

- E.g. 2020 U.S. census test setup with *generous* tabular $\varepsilon = 0.4$
- expectation for individual LAU counts to obtain noise of relative size $\pm 20$, $\pm 50$ and $\pm 100\%$
- analytical estimation (bins) and numerical simulation (lines)
Utility flaws of *unbounded* noise: counts

- Even worse: several counts (e.g. **Total**, **Males**, **Females**) are **distorted** consistently

- E.g. 2020 U.S. census test setup with with **moderate tabular** $\varepsilon = 0.1$

<table>
<thead>
<tr>
<th></th>
<th>2011 census</th>
<th>U.S. setup ($\varepsilon = 0.1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>30</td>
<td>-17</td>
</tr>
<tr>
<td>Male</td>
<td>20</td>
<td>-1</td>
</tr>
<tr>
<td>Female</td>
<td>15</td>
<td>-9</td>
</tr>
</tbody>
</table>
Utility flaws of *unbounded* noise: counts

- Even worse: several counts (e.g. Total, Males, Females) are distorted consistently up or down
- E.g. 2020 U.S. census test setup with moderate tabular $\varepsilon = 0.1$
- still ~20 small LAUs where $\pm 100\%$ would happen (~100 LAUs with $\pm 50\%$)
Utility flaws of *unbounded* noise: ratios

- take very simple *ratio indicator* e.g. share of females: $r := F/T$
  - standard deviation of $r$ as a function of generic noise variance $V$:
    $$\text{sd}_r (V) = \frac{1}{T} \sqrt{V (1 + r^2)}$$
Utility flaws of *unbounded* noise: ratios

- take very simple ratio indicator e.g. share of females: \( r := \frac{F}{T} \)

  ➔ standard deviation of \( r \) as a function of generic noise variance \( V \):

  \[
  \text{sd}_r (V) = \frac{1}{T} \sqrt{V (1 + r^2)}
  \]

- to quantify bound effects, approximate noise effects \( i = i_0 + x_i (i = F, T) \) as

  \[
  r - r_0 = r (\xi_F - \xi_T) + O(\xi^2) \quad \text{with} \quad \xi_i \equiv x_i / i \ll 1
  \]

  ➔ in the presence of a bound \( E \):

  \[
  \max |r - r_0| \simeq \frac{E}{T} (1 + r)
  \]
Utility flaws of *unbounded* noise: ratios

- take very simple ratio indicator e.g. share of females: \( r := \frac{F}{T} \)

  ➔ standard deviation of \( r \) as a function of generic noise variance \( V \):

  \[
  \text{sd}_r (V) = \frac{1}{T} \sqrt{V (1 + r^2)}
  \]

- to quantify bound effects, approximate noise effects \( i = i_0 + x_i (i = F, T) \) as

  \[ r - r_0 = r (\xi_F - \xi_T) + O (\xi_i^2) \text{ with } \xi_i \equiv x_i / i \ll 1 \]

  ➔ in the presence of a bound \( E \):

  \[
  \max |r - r_0| \approx \frac{E}{T} (1 + r)
  \]

- this can be tested numerically with noise samples from CKM (e.g. \( V=3, E=6 \)) and for comparison from unbounded \( \varepsilon \)-DP setup (\( \varepsilon=0.8 \))
Utility flaws of *unbounded* noise: ratios

sanity check on $sd_r$
Utility flaws of *unbounded* noise: ratios

bound effects in \( \max |r-r_0| \)

- *bounded* noise (CK) consistently below model
- *unbounded* noise (DP) consistently above model
- typical size of difference: \(~5 \% \text{ points}\) across bins
- i.e. huge relative diff. for small \( r < 0.1 \) (e.g. minorities)
Additional disclosure risks of *bounded* noise

• Now would you bet all your money on a guess for the **true count** of the …

- … total population?
- … country-born males?
- … total females?
- … total foreign-born?

<table>
<thead>
<tr>
<th>SEX \ POB</th>
<th>Total</th>
<th>Country</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>42</td>
<td>37</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>23</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>

*each count with noise variance* \( V = 1 \)*
*and noise bound* \( E = 2 \)
Additional disclosure risks of *bounded* noise

- Now would you bet all your money on a guess for the true count of the …
  - … total population?
  - ✔️ … country-born males (= 17)
  - … total females?
  - … total foreign-born?

- But **how often** does this happen?

<table>
<thead>
<tr>
<th>SEX \ POB</th>
<th>Total</th>
<th>Country</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>42</td>
<td>37 (= 35+2)</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>23</td>
<td>15 (= 17-2)</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>16 (= 18-2)</td>
<td>3</td>
</tr>
</tbody>
</table>

Each count with noise variance \(V = 1\) and noise bound \(E = 2\)
Additional disclosure risks of *bounded* noise

- linear constraints in breakdowns – e.g. dichotomous $\text{SEX} = \{F,M,T\}$:

  $$\text{expectation} \ (F + M - T) = 0 \ \Rightarrow \ \text{bound estimator} \ \widehat{E} = \left\lfloor \frac{F + M - T}{3} \right\rfloor$$
Additional disclosure risks of *bounded* noise

- linear constraints in breakdowns – e.g. dichotomous SEX = \{F, M, T\}:

  \[ (F + M - T) = 0 \Rightarrow \text{bound estimator} \quad \hat{E} = \left\lfloor \frac{F + M - T}{3} \right\rfloor \]

  prob. to reveal \( E \) from a single 3-tuple: 
  \[ p_1 := \Pr[|F + M - T| > 3(E - 1)] \]

  \( p_1 \) fixed by noise distribution (e.g. CKM pars. \( V \) and \( E \))
Additional disclosure risks of **bounded** noise

- **linear constraints** in breakdowns – e.g. dichotomous $\text{SEX} = \{F, M, T\}$:

  
  
  \[
  \text{expectation} \ (F + M - T) = 0 \Rightarrow \text{bound estimator} \ \widehat{E} = \left\lfloor \frac{F + M - T}{3} \right\rfloor
  \]

  

  prob. to reveal $E$ from a single 3-tuple: $p_1 := \Pr[|F + M - T| > 3(E - 1)]$

  $\Rightarrow p_1$ fixed by noise distribution (e.g. CKM pars. $V$ and $E$)

- **number of 3-tuples needed to disclose $E$ at c.l. $\alpha$**:

  
  \[
  m = \left\lfloor \frac{\log(1 - \alpha)}{\log(1 - p_1)} \right\rfloor
  \]

  $\Rightarrow$ **available** $m$ fixed by table output
Knowing the full output, the risk can be quantified systematically – e.g. for the 2021 EU census output:

$m$: number of 3-tuples needed in output to get ca. one $E$-disclosive noise pattern

black boxes showing where $m$ exceeds the number of available 3-tuples for Malta (dashed) and Germany (solid)

Vanilla CKM from SDCTools on GitHub
Conclusions

• in noisy approaches to confidentiality, whether the noise is *bounded* or *unbounded* is a key question with consequences for both utility and disclosure risks
Conclusions

• in noisy approaches to confidentiality, whether the noise is *bounded* or *unbounded* is a key question with consequences for both utility and disclosure risks – shown today:

• utility – *unbounded* noise cannot guarantee useful outputs on *all* small areas in a large output programme (e.g. EU census LAU data)

  ➔ holds for raw counts and more pronounced for shares/ratios, even with moderate noise variance (e.g. $V \sim 3$)
Conclusions

• in noisy approaches to confidentiality, whether the noise is *bounded* or *unbounded* is a key question with consequences for both *utility* and disclosure *risks* – shown today:

• **utility** – *unbounded* noise cannot guarantee useful outputs on *all* small areas in a large output programme (e.g. EU census LAU data)

  ➔ holds for raw counts and more pronounced for shares/ratios, even with moderate noise variance (e.g. $V \sim 3$)

• **risks** – *bounded* noise is additionally vulnerable to constraint exploits

  ➔ risk can be controlled by tuning noise to output complexity, with moderate noise parameters ($V \sim 2$, $E \sim 5$)
Thank you

© European Union 2023

Unless otherwise noted the reuse of this presentation is authorised under the CC BY 4.0 license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.

Slide XX: map section, source: screenshot from OpenStreetMap; Slide XX: view of Cidamón, source: photo by Bigsus from Wikipedia