Methodology of Advanced Information Analysis (AIA) to data treatment and analysis

T. Ocelka¹, L. Pavliska², P. Kovanic³, J. Rownicka⁴, S. Veronesi¹

¹Institute Environmental Technologies, VSB TU Ostrava, Czech Republic
²E&H services, Inc., Prague, Czech Republic
³Retired Scientist, Prague, Czech Republic
⁴Medical University, Sosnowiec, Poland

Special acknowledgement:
- P. Kovanic – GA founder
- V. Procházka – FNO applications
- L. Pavliska – R- project libraries - development
Why Advanced Information Analysis?
1. If you stacked a pile of CD-ROMs on top of one another until you’d reached the current global storage capacity for digital information – about 295 exabytes – it would stretch 80,000 km beyond the moon.

Source: http://www.kurtosys.com
Data Extremism

2. Every hour, enough information is consumed by internet traffic to fill 7 million DVDs. Side by side, they’d scale Mount Everest 95 times.

3. 247 billion e-mail messages are sent each day... up to 80% of them are spam.
Data Extremism

4. By 2020, IT departments will be looking after 10 x more servers, 50 x more data and 75 x more files.

5. We can expect a 40-60 per cent projected annual growth in the volume of data generated, while media intensive sectors, including financial services, will see year on year data growth rates of over 120 per cent.

Source: http://www.kurtosys.com
Data Extremism

• 6. The world’s 500,000+ data centres are large enough to fill 5,955 football fields.
• 7. 75% of digital information is generated by individuals, whilst enterprises have liability for 80% of digital data at some point in its life.
• 8. There are nearly as many bits of information in the digital universe as there are stars in our actual universe.
• 9. Investment in digital enterprises has increased 50% since 2005.

Source: http://www.kurtosys.com
Data Extremism

10. There are 30 billion pieces of content shared on Facebook every day.

11. In 2010, 28% of the digital universe required some level of security... not all of it had the level of security it required....

12. People wishing each other Happy New Year drove a 500% surge in smartphone data within just one year, according to 3UK whose customers used a whopping 80 terabytes (TB) on the 31st December 2011, compared to just 14 TBs on the same day in 2010.

Source: http://www.kurtosys.com
Each YEAR – data increase

x2
The GOAL
Practical insight into DA

Studies and data characters
Introduction to novel approach
Elementary outputs
AIA – HOW TO
Conclusions
Analytical background - NRL for POPs

- Accreditation EN ISO/IEC 17 025
- Realization of accreditation SNAS
- Recognition ILAC, EA, IAF
- Sampling and diagnostics
- Circular analyses
  - CR (indicators of PCBs)
  - International (toxic PCBs)
- Authorization NRL for POPs
  - Min. of Agriculture
  - Min. of Health
- NRL involvements
  - Community Reference Laboratory for Dioxins and PCBs in Feed and Food in Freiburg

---

**GC-HRMS (POPs)**
- MAT 95XP
- from 2003

**LC-MS/MS** (pharmaceuticals, pesticides)
- ThermoFinnigan
- from 2006

**GC-MS/MS (ion traps)**
- GCQ, Polaris
How our data look like?

- Large variability
- Low scaled series (low repeatability 2 – 10)
- High uncertainty (trace detection)
- Apriori assumption on NORMALITY data = Science Fiction; squewed data
- Data are polymodal
- LOD values
<table>
<thead>
<tr>
<th>Method</th>
<th>SOP 35</th>
<th>SOP 35</th>
<th>SOP 35</th>
</tr>
</thead>
<tbody>
<tr>
<td>UHF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exposure</th>
<th>E189</th>
<th>E260</th>
<th>E260</th>
<th>E260</th>
<th>E260</th>
<th>E260</th>
</tr>
</thead>
<tbody>
<tr>
<td>E189</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>E189</th>
<th>E260</th>
<th>E260</th>
<th>E260</th>
<th>E260</th>
<th>E260</th>
</tr>
</thead>
<tbody>
<tr>
<td>E189</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E260</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| E189   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |

| E189   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |

| E189   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |

| E189   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |

| E189   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |

| E189   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |

| E189   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |

| E189   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
| E260   |      |      |      |      |      |      |
Present status – the paradigm of statistics

- Excellent methodology
- Many applications
- Strong SW support
- Apriori distribution
- Normality assumption
- Small data samples (N>10 is accepted, e.g. Horn method)
- Central Limit Theorem – large data set for normality assumption
- Momentum parameters
Novel approach requirements

Strong theoretical background
Its laws respecting fundamentals of our applications = natural laws
One universal determined way
Availability – the tool / cost
  • R-project / LabKey platform
Feasibility - as easy as 1-2-3

Robust Advanced Information Analysis
IA analysis development

The paradigm
2000 – first environmental application
- Only limited ideas – statistics was considered as a dogma
2001-2003 – other applications
- Environmental, Preventive medicine
Since 2012 – broader spectrum of applications
- Combination health-environmental data (passive sampling)

Tools
- S-Plus (v 6.2)
- Open-source system „R-project“ (since 2008)
- International projects - EU: MAGIC, FOKS, 2-FUN

Present
- Development - CEDAN concept
Two paradigms (Statistics X AIA)

Data

Asp

Trans

MS

Marginal DA

MVDA

Common format

Assumptions
Normality
Variance
Homogeneity

eg. Box-Cox

Method selection
Data/ regression model / estimation method

Data

• Data Warehouse
• The standard (e.g. HL7)
• FileMaker™/LabKey
• ehCloud for sharing
• On-line computing (IT4I)

Marginal DA

MVDA
Practical studies

Passive sampling devices
state of art in AIA

The package
Open Source
The Guide
The Theory
Case studies
Dev. team
Sampling sites

Brno – rivers+ WWTP

Jaworzno – Industrial landfill
## Parameters of distribution

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowb (LB)</td>
<td>the lower bound of the data support</td>
</tr>
<tr>
<td>LSB</td>
<td>the lower bound of the membership interval</td>
</tr>
<tr>
<td>RobMed</td>
<td>quantile of probability 0.5</td>
</tr>
<tr>
<td>USB</td>
<td>the upper bound of the membership interval</td>
</tr>
<tr>
<td>Uppb (UB)</td>
<td>the upper bound of the data support</td>
</tr>
<tr>
<td>RobMean</td>
<td>robust mean value of the data</td>
</tr>
<tr>
<td>STD.DF</td>
<td>root square of the variance</td>
</tr>
</tbody>
</table>
IA - the comparison of SP
AIA – Correlation matrix
The future...

EXPECT SUCCESS

Just Go Ahead
Increasingly, the datasets are so large, and the application programs are so complex, that it is much more economical to move the end-user’s programs to the data and only communicate questions and answers rather than moving the source data and its applications to the user’s local system.

Gray, et.al., 2005)
Data are too expensive if not transferred into valid information

Building data store
Design and optimalisation of proper interface (HL 7 like)
Allow data from multi-sources to be integrated commonly indexed

- Definition of critical infrastructure
- Support to PM, Workflow, and strategic RHC-IS
- On-line transaction support – IT4I interface
- Administration

- Design and development of robust DA tool
- Optimisation to R-project environment
- Transfer to WEB-based environment

- Set-up working groups (IT+EG) for DA by interaction
- Marginal and multivariate data analysis
- Interpretation
- Case-studies store for KH transfer
Conclusions

The process extraction maximum information from data can be fun reality

AIA hides strong application potential, as verified in economical, environmental and medical studies

Open-source environment R-project® and LabKey® for sharing of studies/data


Ready To Go team to share experience
Thank you for your attention!