Design, evolution and optimization of monitoring networks using information concepts

Axel Osses (Universidad de Chile)
in coll. with L. Gallardo, T. Faúndez, A. Henríquez, M. Díaz

DIM - Departamento de Ingeniería Matemática
www.dim.uchile.cl

CMM - Center for Mathematical Modeling (UMI 2601 CNRS)
www.cmm.uchile.cl

(CR)$^2$ - Center for Climate and Resilience Research
www.cr2.cl

Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago.

Workshop Big Data Environment, Nov 10th-13th 2015
Air-quality monitoring network analysis

- Monitoring network: multi-objective (quality standards, control, curbing measures, impacts on health, ecosystems, climate, etc.).

Monitoring network design/analysis

- where to place new stations of the network?
- which stations could be removed?
- optimal geographical distribution? which criteria?

- An increasing research oriented towards network design\(^1\).

- We introduced some statistical and variational indicators for network design derived from information theory\(^2\).

---


\(^2\) Boltzmann/Gibbs 1870s, Shannon 1948, Kullback 1959
Santiago’s air quality network

Santiago’s air quality network

Figure: PM10 and O3 measurements
Santiago’s air quality network

Figure: PM10 “dosis”
Santiago’s air quality network

Figure: O3 “dosis”
I.- Intro

Data

Data base

1997-2008, 7 stations

<table>
<thead>
<tr>
<th></th>
<th>normal</th>
<th>log-normal</th>
<th>gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>S</td>
<td>W</td>
</tr>
<tr>
<td>CO</td>
<td>26.9</td>
<td>24.1</td>
<td>18.1</td>
</tr>
<tr>
<td>O₃</td>
<td>9.76</td>
<td>10.5</td>
<td>8.99</td>
</tr>
<tr>
<td>PM₁₀</td>
<td>10.5</td>
<td>6.46</td>
<td>8.96</td>
</tr>
<tr>
<td>SO₂</td>
<td>37.1</td>
<td>42.7</td>
<td>26.0</td>
</tr>
</tbody>
</table>

2009-2010, 11 stations

<table>
<thead>
<tr>
<th></th>
<th>normal</th>
<th>log-normal</th>
<th>gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>S</td>
<td>W</td>
</tr>
<tr>
<td>PM₁₀</td>
<td>16.3</td>
<td>7.63</td>
<td>8.55</td>
</tr>
<tr>
<td>PM₂₅</td>
<td>9.84</td>
<td>4.69</td>
<td>9.59</td>
</tr>
<tr>
<td>O₃</td>
<td>10.9</td>
<td>12.6</td>
<td>6.79</td>
</tr>
</tbody>
</table>

Table: Relative quadratic error (%) for different data fitting.

Figure: Example of some statistical fitting at 2 stations.
Evolution of the network

Figure: Evolution of Santiago’s air monitoring sites and urban-rural limit.
II.- Statistical indicators linked to “information”³

Quality indicators:

- mutual information or “specificity index”:
  
  how difficult is to reproduce measurements of i-th station from the complementary measurements on the network?

- information gain or “representativity index”:
  
  total information gain.

- information gaps associated to the evolution of a network;

They are introduced based on the concept of relative information or “divergence” by Kullback and Liebler.

We use them to analyze 14 years of Santiago’s network public data (1997-2010).

Design, evolution and optimization of monitoring networks using information concepts

II.- Statistical Analysis

Concepts

Basis: Kullback-Liebler divergence between distributions

Kullback-Liebler divergence of $q_X$ w.r.t. $p_X$

$$KL(p_X \| q_X) = \int p_X(x) \ln \frac{p_X(x)}{q_X(x)} dx,$$


Normal case: $p_X \sim \mathcal{N}(\mu_0, \Sigma_0)$, $q_X \sim \mathcal{N}(\mu_1, \Sigma_1)$

$$KL = \frac{1}{2} \left( \text{tr}(\Sigma_1^{-1}\Sigma_0) - nm - \ln \left| \frac{\Sigma_0}{\Sigma_1} \right| + \Sigma_1^{-1} (\mu_0 - \mu_1)^2 \right).$$

$\Sigma$: covariance matrix, $\text{tr}$: trace, $| \cdot |$: determinant.

$KL \geq 0$ vanishes only if $p_X = q_X$ but is non symmetric.
Mutual Information and Specificity index

Mutual info between $i$th-station and other stations (complement)

\[
I^i_M = KL(p_X \| p_{X_i} p_{X^c_i}) = - \frac{1}{2} \ln \frac{|\Sigma_X|}{|\Sigma_{X^c_i}| |\Sigma_{X_i}|}.
\]

- $p_{X_i}, p_{X^c_i}$: marginal densities, $p_X$: joint density.
- $p_{X_i} = \mathcal{N}(\mu_{X_i}, \Sigma_{X_i}), p_{X^c_i} = \mathcal{N}(\mu_{X^c_i}, \Sigma_{X^c_i}), p_X = \mathcal{N}(\mu, \Sigma_X)$

Specificity index

\[
s_i = 1 - \frac{I^i_M}{\max_j I^j_M} \quad i = 1, \ldots, n.
\]

→ how difficult is to reproduce measurements of $i$-th station from the complementary measurements on the network?
Design, evolution and optimization of monitoring networks using information concepts

II.- Statistical Analysis

Concepts

Information Gain and Representativity index

**Information gain by measurements of \( i \)-th station**

\[
I^i_G = \text{KL}(p_X \| q_{X^c_i}) = \frac{1}{2} \left( \text{tr}(B_i^{-1} \Sigma_i) - m \ln \frac{|\Sigma_X|}{|\Sigma_{X^c_i}||B_i|} + B_i^{-1}(\mu_X - \mu_b)^2 \right)
\]

- \( q_{X^c_i}, p_X \): situations before and after \( i \)-th measurements.
- \( q_{X^c_i} \sim \mathcal{N}(\mu'_i, \Sigma'_i), \mu'_i = (\mu_{b_i}, \mu_{X^c_i}), \Sigma'_i = \text{diag}(B_i; \Sigma_{X^c_i}) \)
- \( \mu_{b_i}, B_i \): a priori background mean and covariance of \( i \)th-station.

**representativity index of the \( i \)-th station**

\[
r_i = \frac{I^i_G}{\max_j I^j_G}, \quad i = 1, \ldots, n.
\]

→ relative information gain. We can also compute the information gain \( I^K_G \) associated to a subset of stations \( K \subset \{1, \ldots, n\} \).
Information gaps and evolution of total information

Information gap from $K_1$ to $K_2$

$$\Delta I^{K_1,K_2} = \text{KL}(p_X || q_{K_1}) - \text{KL}(p_X || q_{K_2}) = I_{G}^{K_1} - I_{G}^{K_2}.$$ 

can be positive or negative and $\Delta I^{K_1,K_2} + \Delta I^{K_2,K_3} = \Delta I^{K_1,K_3}$.

**Figure:** Evolution of total information
II.- Statistical Analysis

Concepts

Normalized information distance and clustering

**Mutual information between stations $i$ and $j$**

$$I_{ij}^M = \text{KL}(p_{X_i, X_j} \| p_{X_i} p_{X_j}).$$

- $p_{X_i, X_j}$: joint, $p_{X_i}$, $p_{X_j}$: marginals.

**Normalized information distance between stations $i$ and $j$**

$$d_{ij} = 1 - \frac{I_{ij}^M}{\max(H_i, H_j)},$$

- $H_i = -\sum_x p_{X_i}(x) \ln p_{X_i}(x)$: Shannon entropy of measurements $X_i$.

This distance is zero if and only if $p_{X_i}$ and $p_{X_j}$ are independent (this is not the case for the Pearson’s correlation coefficient).
Design, evolution and optimization of monitoring networks using information concepts

II.- Statistical Analysis

Removing stations

**Figure:** Specificity (top) and representativity (bottom) indexes (simultaneously) for CO, O₃, PM₁₀ and SO₂ for hourly data for the period 1997–2008 in summer, winter and all seasons. Larger circle → larger index.
Where to add a new station?

**Figure:** Left: simulated Barnes interpolation (log [PM$_{10}$], lighter=higher). Right: at each point, information gain obtained if we add a new station with interpolated values.
II.- Statistical Analysis

Evolution

Evolution analysis

Figure: Simulated evolution of total information content, considering PM$_{2.5}$ measurements 2009-2010.

Figure: The same as before using a interpolated priori information (Barnes interpolation).
Clustering analysis

PM$_{2.5}$

**Figure:** Hierarchical clustering using the normalized information distance (left column) compared with the Pearson correlation function (right column) (2009-2010).
II.- Variational indicators linked to “information” \(^4\)

**Quality indicators:**
- Precision gain
- Total information gain
- Degrees of freedom

They are introduced in the data assimilation framework.

We use them, *weighted* by some design criteria, to reduce, extend and optimize the air-quality monitoring network of Santiago.

Given a linear tracer, meteorology, the sensitivity matrix $H$ store the impact of unit emissions at sites $X$ in measurements sites $Y$:

$$Y = H X$$

The best estimator of true emissions, is the unique solution of:

$$\min_X \frac{1}{2} \| H X - Y_o \|_R^2 + \frac{1}{2} \| X - X_b \|_B^{-1}$$

$Y_o$: $m$-dimensional measurement vector with covariance $R$.
$X_b$: background estimation (best guess) with covariance $B$.

**Analysis: best estimator of emissions and its covariance**

$$X_a = X_b + \Sigma_a^{-1} (HX_b - Y_0)$$

$$\Sigma_a = (B^{-1} + H^t R^{-1} H)^{-1}$$
One network = one subsensitivity

Each monitoring network can be characterized by a submatrix $H'$ of the total sensitivity $H$ with associated analysis $X'_a$, $\Sigma'_a$:

**Figure:** Left: network sites in emission grid. Center: selected sites as rows of the total sensitivity. Right: reduced sensitivity matrix.
Precision gain of a network

The *precision gain* is obtained by subtracting the total precision after and before the observations of the network are assimilated:

$$\Delta pr(H') = \text{Tr} \left( \Sigma_a^{-1} \right) - \text{Tr} \left( B^{-1} \right), \quad H' \leftrightarrow \text{network}$$
Total information gain of a network

The *information gain* of the network is obtained by subtracting the total information after and before the observations of the network are assimilated:

\[
\Delta I(H') = \frac{1}{2} \ln |B| - \frac{1}{2} \ln |\Sigma'|, \quad H' \leftrightarrow \text{network}
\]
Degrees of freedom of a network

The degrees of freedom represents the number of states (in the $n$-dimensional emission space) that can be effectively retrieved from the observations of the given network:

$$d.f.(H') = n - \text{Tr}(B^{-1}\Sigma_a'), \quad H' \leftrightarrow \text{network}$$

- limit cases: no knowledge $g.l. = 0$, perfect knowledge $g.l. = n$.
- The degrees of freedom corresponds to the trace of the so called influence matrix $A$:

$$A = R^{-\frac{1}{2}} H' \Sigma_a H'^t R^{-\frac{1}{2}}.$$
II.- Variational analysis

Degrees of freedom

<table>
<thead>
<tr>
<th>Quality indicator</th>
<th>Q</th>
<th>Definition</th>
<th>$f \left( \frac{\lambda}{\mu} \right)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision gain</td>
<td>$\Delta pr$</td>
<td>$\text{Tr}(\Sigma_a^{-1}) - \text{Tr}(B^{-1})$</td>
<td>$\frac{1}{\sigma_b^2} \frac{\lambda^2}{\mu^2}$</td>
</tr>
<tr>
<td>Information gain</td>
<td>$\Delta I$</td>
<td>$\frac{1}{2} \ln</td>
<td>\Sigma_a^{-1}</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>$d.f.$</td>
<td>$n - \text{Tr}(B^{-1}\Sigma_a)$</td>
<td>$1 - \left(1 + \frac{\lambda^2}{\mu^2}\right)^{-1}$</td>
</tr>
</tbody>
</table>

Table 1. Summary of the main quality indicators or metrics for an air quality network. See text for details, and Figure 2 for illustration.
Weights

We apply a weight $\sqrt{\pi_j^\beta}$ to each emission grid point $j$ ($\beta$: modulation parameter), for example:

- population density
- health risk
- feasibility costs

So we replace $H$ by

Weighted sensitivity

\[ H'_\beta = H' \Pi_\beta \]

before computing the quality indicators, where:

\[ \Pi_\beta = \text{diag} \left( \sqrt{\pi_1^\beta}, \ldots, \sqrt{\pi_n^\beta} \right). \]
II.- Variational analysis

Weighted sensitivity

Population density weights

Figure: Weighting functions applied. Upper panel: log of population density (hab/km²). Lower panel: log of normalized CO summer emission fluxes (molkm⁻²hr⁻¹). White contour: urban-rural limit in 2010. White circles: location of monitoring stations in 2009.
Removing stations

Figure: Total information gain without (left) or with (right) population density weight. Remove stations with smallest circles.
Design, evolution and optimization of monitoring networks using information concepts

Adding stations

**Figure:** White squares: potential location of new stations coinciding with local maxima of information gain (in percentage w.r.t. basal network).
Optimal network design

Figure: Optimal networks and wind patterns.
Optimal placement and evolution

Figure: Real v/s Optimal evolution: 4, 8, 11, 14 stations. Squares: new stations.
Optimal placement: 4 stations

**Figure:** Network search: maximizing total information ($- \min_H(-\Delta I)$) with weights.
Summary

- Indicators (both statistical/variational) can be used concurrently to analyse/design an observational network.

  - **Statistical indicators.** **Pros:** simple for remove/analyze, use real measurements. **Cons:** do not include dispersion models, adding stations involves hard interpolation (kriging, variograms) not physically consistent.

  - **Variational indicators.** **Pros:** include dispersion models so analysis (add/remove/optimize) is consistent with known emission and circulation patterns, weight criteria allowed. **Cons:** measurements are not directly used (but could be indirectly used via weights and/or data assimilation). Time consuming modeling.
Many thanks!