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].

[1] Perez-Abreu 1996, Saunier 2009, Ruiz 2010, Bocquet 2011, Ruiz 2012, Zidek 2010

[2] Boltzmann/Gibbs 1870s, Shannon 1948, Kullback 1959

Santiago's air quality network

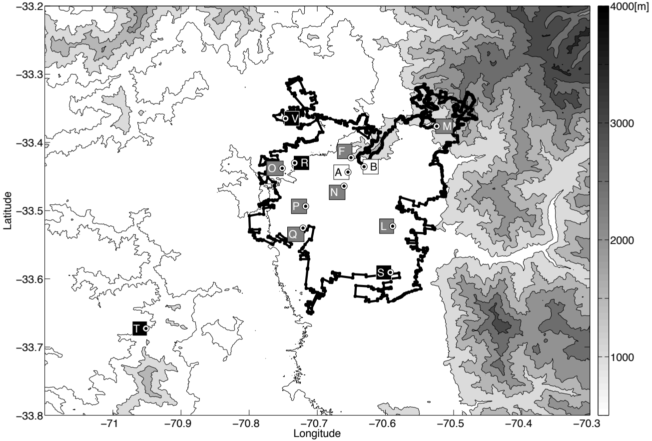


Figure: Stations **A**: Gotuzzo, **B**: Providencia (not used, in white); **F**: Independencia, **L**: La Florida, **M**: Las Condes, **N**: Parque O'Higgins, **O**: Pudahuel, **P**: Cerrillos, **Q**: El Bosque (1997-2008, in gray), **R**: Cerro Navia, **S**: Puente Alto, **T**: Talagante, **V**: Quilicura (2009-2010, in black)

Santiago’s air quality network

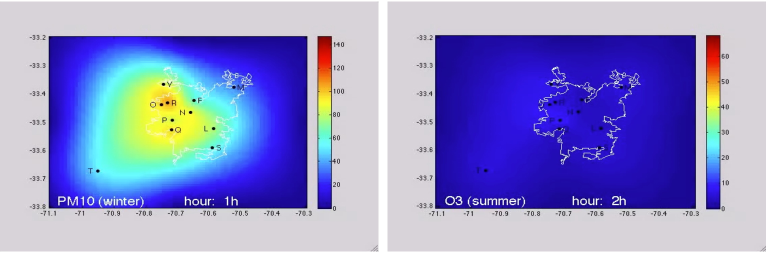


Figure: PM10 and O3 measurements

Santiago's air quality network

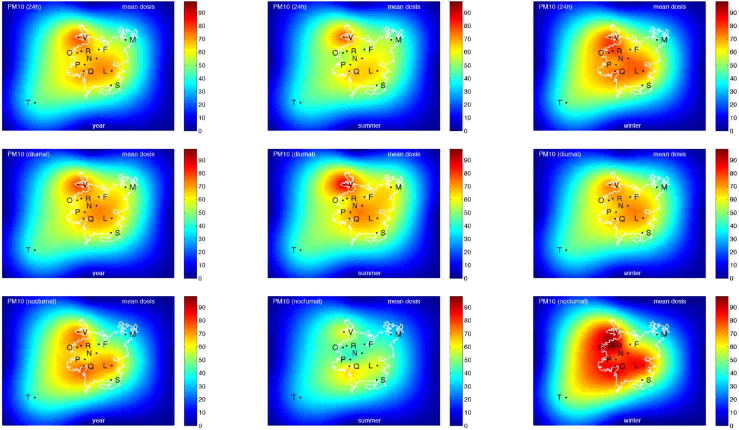


Figure: PM10 "dosis"

Santiago's air quality network

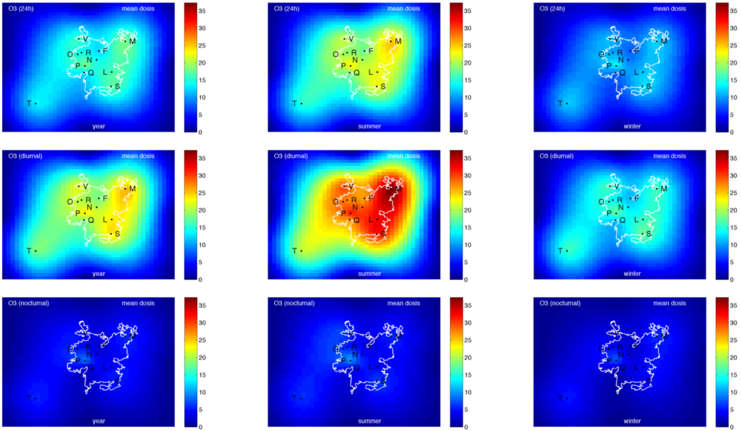


Figure: O3 "dosis"

Data base

1997-2008, 7 stations

	normal			log-normal			gamma		
	All	S	W	All	S	W	All	S	W
CO	26.9	24.1	18.1	14.9	23.9	10.9	5.17	8.07	4.66
O_3	9.76	10.5	8.99	11.1	19.7	8.87	3.93	10.5	2.06
PM_{10}	10.5	6.46	8.96	1.87	1.50	3.94	0.63	0.41	1.17
SO_2	37.1	42.7	26.0	41.1	40.2	30.5	21.0	24.2	12.7

2009-2010, 11 stations

	normal			log-normal			gamma		
	All	S	W	All	S	W	All	S	W
PM_{10}	16.3	7.63	8.55	1.78	1.08	3.38	2.25	0.70	1.06
$PM_{2.5}$	9.84	4.69	9.59	1.36	1.49	2.29	0.71	0.48	0.95
O_3	10.9	12.6	6.79	9.81	16.9	6.89	3.85	9.78	3.30

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

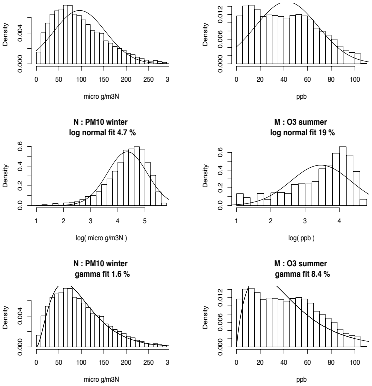


Figure: Example of some statistical fitting at 2 stations.

Evolution of the network

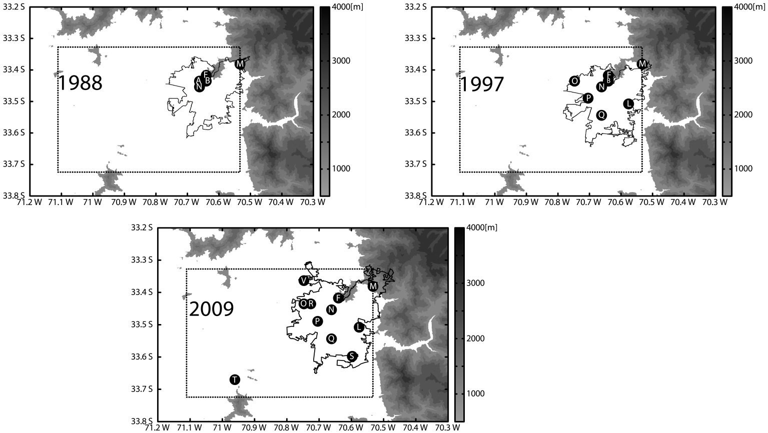


Figure: Evolution of Santiago's air monitoring sites and urban-rural limit.

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).

[3] A. Osses, L. Gallardo, T. Faúndez, Tellus B, 65, 2013

Basis: Kullback-Liebler divergence between distributions

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

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

X : multivariate vector of measurements. n : stations, m : species.
p_X : reference distribution q_X : perturbed distribution.

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

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

Σ : covariance matrix, tr : trace, $|\cdot|$: determinant.
$\mathrm{KL} \geq 0$ 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)

$$I_M^i = \mathrm{KL}(p_X \| p_{X_i} p_{X_i^c}) = -\frac{1}{2} \ln \frac{|\Sigma_X|}{|\Sigma_{X_i^c}|\,|\Sigma_{X_i}|}.$$

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

□ specificity index

$$s_i = 1 - \frac{I_M^i}{\max_j I_M^j} \qquad i = 1, \ \ldots, \ n.$$

$\rightarrow$ how difficult is to reproduce measurements of *i*-th station from the complementary measurements on the network?

Information Gain and Representativity index

Information gain by measurements of *i*-th station

$$I_G^i = \mathrm{KL}(p_X \| q_{X_i^c}) = \frac{1}{2}\left(\mathrm{tr}(B_i^{-1}\Sigma_{X_i}) - m - \ln\frac{|\Sigma_X|}{|\Sigma_{X_i^c}|\,|B_i|} + B_i^{-1}(\mu_{X_i} - \mu_{b_i})^2\right)$$

- $q_{X_i^c}$, p_X: situations before and after i-*th* measurements.
- $q_{X_i^c} \sim \mathcal{N}(\mu_i', \Sigma_i')$, $\mu_i' = (\mu_{b_i}, \mu_{X_i^c})$, $\Sigma_i' = \mathrm{diag}(B_i; \Sigma_{X_i^c})$
- μ_{b_i}, B_i: a priori background mean and covariance of *i*th-station.

□ representativity index of the *i*-th station

$$r_i = \frac{I_G^i}{\max_j I_G^j} \qquad i = 1, \ldots, n.$$

$\rightarrow$ relative information gain. We can also compute the information gain I_G^K 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

$$\Delta I^{K_1,K_2} = \mathrm{KL}(p_X \| q_{K_1}) - \mathrm{KL}(p_X \| q_{K_2}) = I_G^{K_1^c} - I_G^{K_2^c}.$$

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

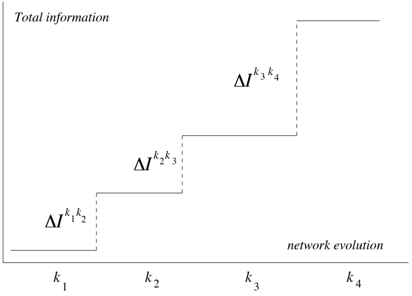


Figure: Evolution of total information

Normalized information distance and clustering

Mutual information between stations i and j

$$I_M^{ij} = \mathrm{KL}(p_{X_i,X_j} \| p_{X_i} p_{X_j}).$$

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

☐ normalized information distance between stations i and j

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

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

Removing stations

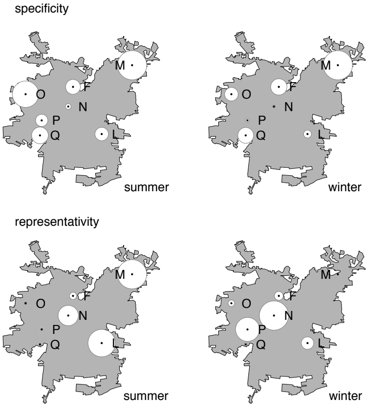
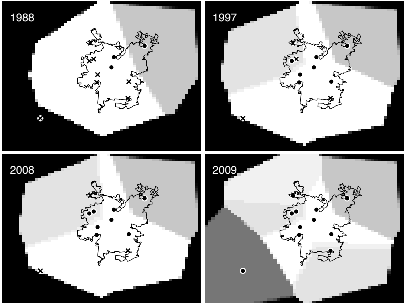


Figure: Specificity (top) and representativity (bottom) indexes (simultaneously) for CO, O_3, PM_{10} and SO_2 for hourly data for the period 1997–2008 in summer, winter and all seasons. Larger circle → larger index.

Where to add a new station?



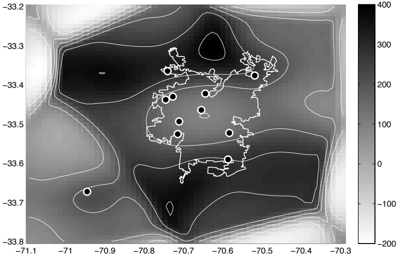


Figure: Left: simulated Barnes interpolation (log [PM_{10}], lighter=higher). Right: at each point, information gain obtained if we add a new station with interpolated values.

Evolution analysis

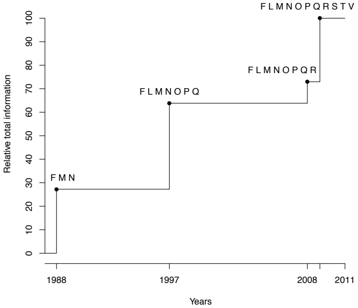


Figure: Simulated evolution of total information content, considering $PM_{2.5}$ measurements 2009-2010.

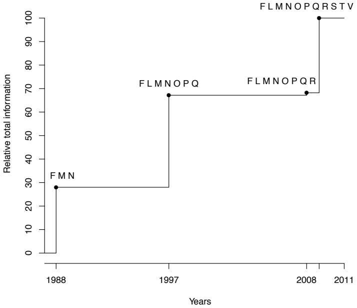


Figure: The same as before using a interpolated priori information (Barnes interpolation).

Clustering analysis

PM$_{2.5}$

O$_3$

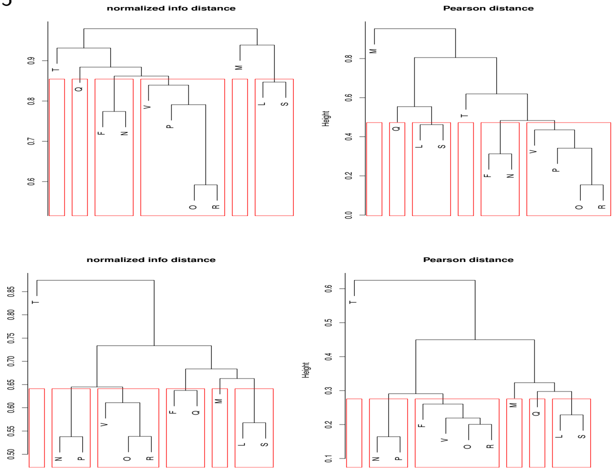


Figure: Hierarchical clustering using the normalized information distance (left column) compared with the Pearson correlation function (right column) (2009-2010).

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.

[4] A. Henríquez, A. Osses, L. Gallardo, M. Díaz, to appear in Tellus B 2015

Data assimiliation framework

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

$$Y = H\,X$$

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

$$\min_X \frac{1}{2}\|HX - Y_o\|^2_{R^{-1}} + \frac{1}{2}\|X - X_b\|_{B^{-1}}$$

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

Analysis: best estimator of emmissions 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, Σ'_a:

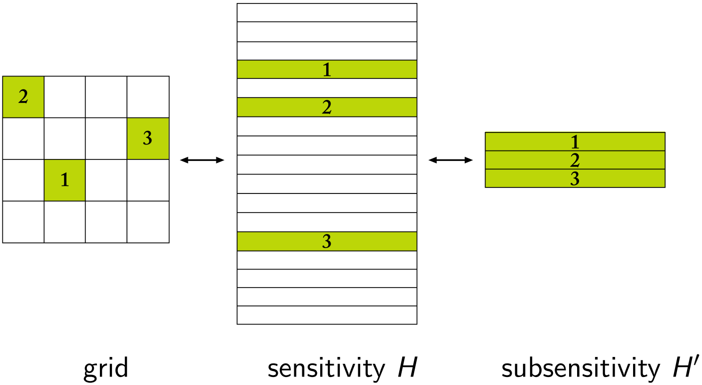


Figure: Left: network sites in emmission grid. Center: selected sites as rows of the total sensitivity. Right: reduced sensitivity matrix.

Precision gain of a network

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

precision gain

$$\Delta pr(H') = \text{Tr}\left(\Sigma_a'^{-1}\right) - \text{Tr}\left(B^{-1}\right), \qquad H' \leftrightarrow \text{network}$$

Total information gain of a network

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

☐ total information gain

$$\Delta I(H') = \frac{1}{2}\ln|B| - \frac{1}{2}\ln|\Sigma'_a|, \qquad H' \leftrightarrow \text{network}$$

Degrees of freedom of a network

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

☐ degrees of freedom

$$d.f.(H') = n - \mathrm{Tr}(B^{-1}\Sigma'_a), \qquad H' \leftrightarrow \text{network}$$

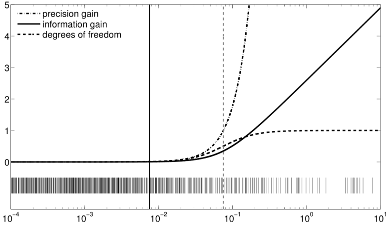
- limit cases: no knowledge $g.l. = 0$, perfect knowledge $g.l. = n$.

- The degrees of freedom corresponds to the trace of the so called *influence matrix A*:

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

Quality indicator	Q	Definition	$f\left(\frac{\lambda}{\mu}\right)$
Precision gain	Δpr	$\mathrm{Tr}\left(\Sigma_a^{-1}\right) - \mathrm{Tr}\left(B^{-1}\right)$	$\frac{1}{\sigma_b^2}\frac{\lambda^2}{\mu^2}$
Information gain	ΔI	$\frac{1}{2}\ \ln\lvert\Sigma_a^{-1}\rvert - \frac{1}{2}\ \ln\lvert B^{-1}\rvert$	$\frac{1}{2}\ln\left(1+\frac{\lambda^2}{\mu^2}\right)$
Degrees of freedom	$d.f.$	$n - \mathrm{Tr}(B^{-1}\Sigma_a)$	$1-\left(1+\frac{\lambda^2}{\mu^2}\right)^{-1}$

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 (β: modulation parameter), for example:

weights

- population density
- health risk
- feasibility costs

So we replace H by

Weighted sensitivity

$$H'_\beta = H'\Pi_\beta$$

before computing the quality indicators, where:

$$\Pi_\beta = \mathrm{diag}\,(\sqrt{\pi_1^\beta}, \ldots, \sqrt{\pi_n^\beta}).$$

Population density weights

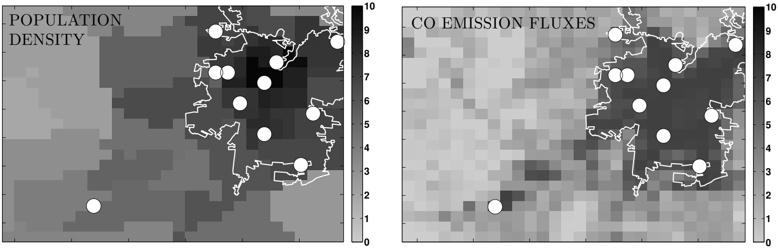
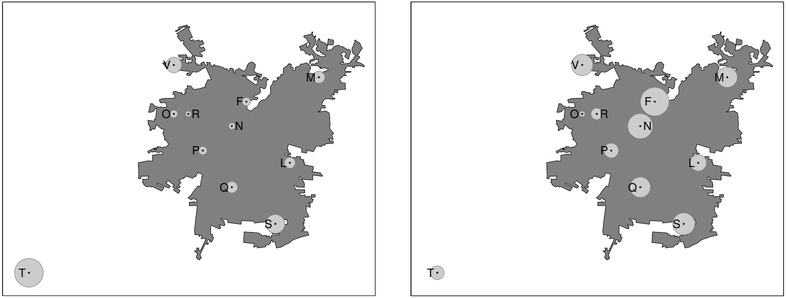


Figure: Weighting functions applied. Upper panel: log of population density (hab/km^2). Lower panel: log of normalized CO summer emission fluxes ($molkm^{-2}hr^{-1}$). White contour: urban-rural limit in 2010. White circles: location of monitoring stations in 2009.

Removing stations



$\beta = 0$ $\beta = 0.7$

Figure: Total information gain without (left) or with (right) population density weight. Remove stations with smallest circles.

Adding stations

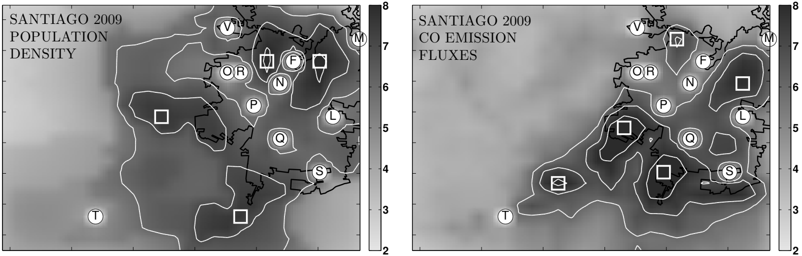


Figure: White squares: potential location of new stations coinciding with local maxima of information gain (in percentage w.r.t. basal network).

Optimal network design

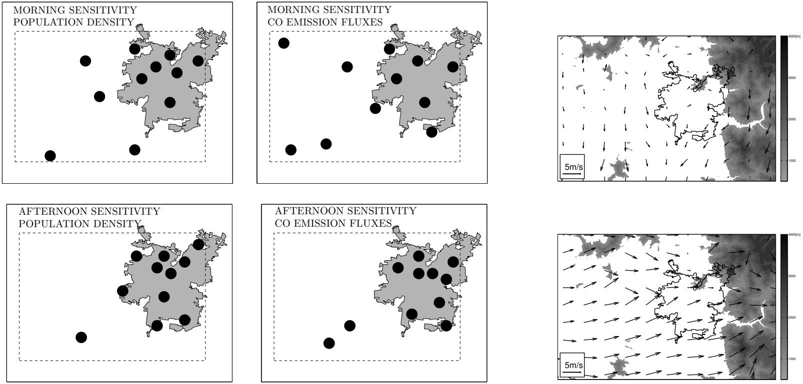


Figure: Optimal networks and wind patterns.

Optimal placement and evolution

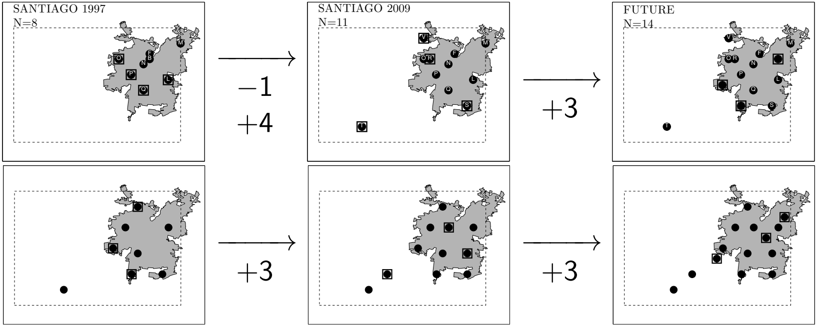


Figure: Real v/s Optimal evolution: 4, 8, 11, 14 stations. Squares: new stations.

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, weigth criteria allowed. **Cons**: measurements are not directly used (but could be indirectly used via weights and/or data assimilation). Time consuming modeling.

Many thanks!