CONSORZIO LaMMA

COOperative Satellite navigation for MEteo-marine MOdelling and Services

COSMEMOS

"Many a little makes a mickle"

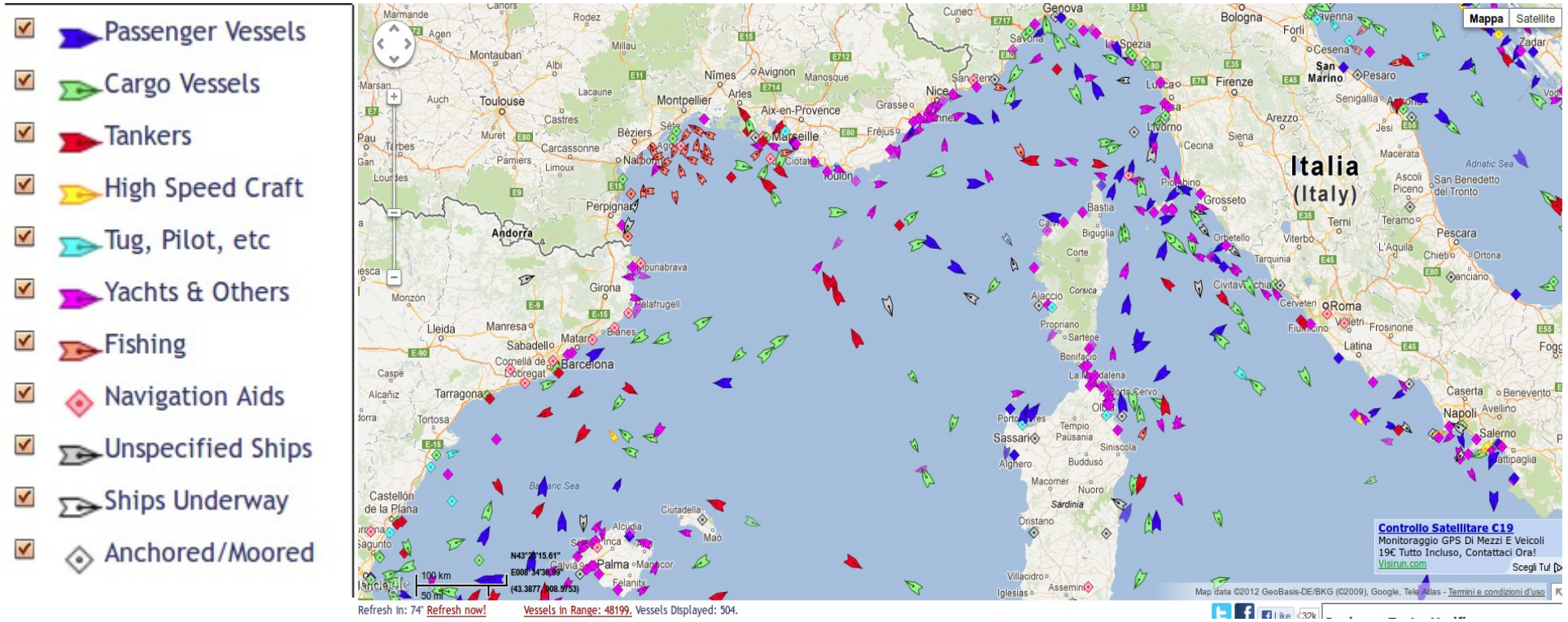
the cooperative data collection for data assimilation,
i.e. how to accumulate small pieces of information without wasting them

Riccardo Benedetti

Meteo-marine data from a (potential) collaborative fleet

From **AIS** data visualized by Marinetraffic.com:
All vessels, on 28 August 2012 at about 16:00 (local time)
Total number of vessels in the scene: **504**

MarineTraffic.com



On board sensors for: temperature (air and sea water), pressure, wind speed and direction, position (GNSS), humidity …

On board observations and data assimilation for NWP

Cooperative data features	Data assimilation perspective
➢ Nearly continuous sampling	Not a problem
➢ Sea coverage	Added value
➢ Large amount of data	Not a problem
➢ Observations from heterogeneous not calibrated instruments	Problem: a quantitative (not so large) observation error is required
➢ Fragmentary character of information	Problem: combination of many pieces of information from different sources

No information can be wasted, all the information has to be elaborated

"Many a little makes a mickle": from money to information

The old Scottish proverb is clear when applied to money (additive quantity):

It should be also valid for the pieces of information: the evidence from a large number of observations is expected to be stronger than that from any single experiment.

"*Molti indizi fanno una prova*" ("*many clues make a proof*")...

But how to combine evidence from different sources?
Intuition and common sense are not enough...

A naive approach (or fraud) with disastrous consequences

Example - Ship routing: NEW method versus OLD method comparison.

A+B (pooling the data)

	Failures	Successes	Percentage of Success [%]
OLD method	1000	760	43.18 ± 1.18
NEW method	145	187	56.33 ± 2.72

The new method is better without reasonable doubt!

Or not?

=

Experiment A (e.g. on summer)

	Failures	Successes	Percentage of Success [%]
OLD method	800	200	20.00 ± 1.26
NEW method	75	12	13.79 ± 3.70

The old method was better...

+

Experiment B (e.g. on autumn)

	Failures	Successes	Percentage of Success [%]
OLD method	200	560	73.68 ± 1.60
NEW method	70	175	71.43 ± 2.89

The old method was better...

What is going wrong in pooling A and B data?

A naive approach (or fraud) with disastrous consequences

$R_{1:T}=\{R_1, R_2, \ldots R_T\}$ time series of results (failure=0/success=1) for the method under test

$S=$ number of successes in $R_{1:T}$

$S_q=$"the probability of success for the method under test is q"

$I=$prior information (i.e. before the test results are known)

$$p(S_q|R_{1:T}I)=?$$

$$p(S_q|R_{1:T}I)=p(S_q|I)\frac{p(R_{1:T}|S_qI)}{p(R_{1:T}|I)}\propto$$

normalisation constant

$$p(S_q|I)\,p(R_1|S_qI)\,p(R_2|R_1S_qI)\ldots p(R_T|R_1R_2\ldots R_{T-1}S_qI)=$$

$$p(S_q|I)\,p(R_1|S_qI)\,p(R_2|S_qI)\ldots p(R_T|S_qI)$$

prior (e.g. uniform)

$$q^S(1-q)^{T-S}$$

$$p(R_i|S_qI)=\begin{cases} q & \text{if } R_i=1 \\ 1-q & \text{if } R_i=0 \end{cases}$$

Commutative product of T factors, always the same for any given S! But...

001000100000100 111011110101101 ≠ 101110010100101 100101000110101

A B

How to make the analysis sensible to the data sequence

"I beseech you, in the bowels of Christ, think it possible that you may be mistaken" Oliver Cromwell to the synod of the Church of Scotland – 1650

According to the "Cromwell principle" we have to consider the possibility that:

$E =$"An extra factor changing in unknown way the probability q of success is present"

$p(S_q|R_{1:t}EI) = p(S_q|I)$ (uniform distribution)

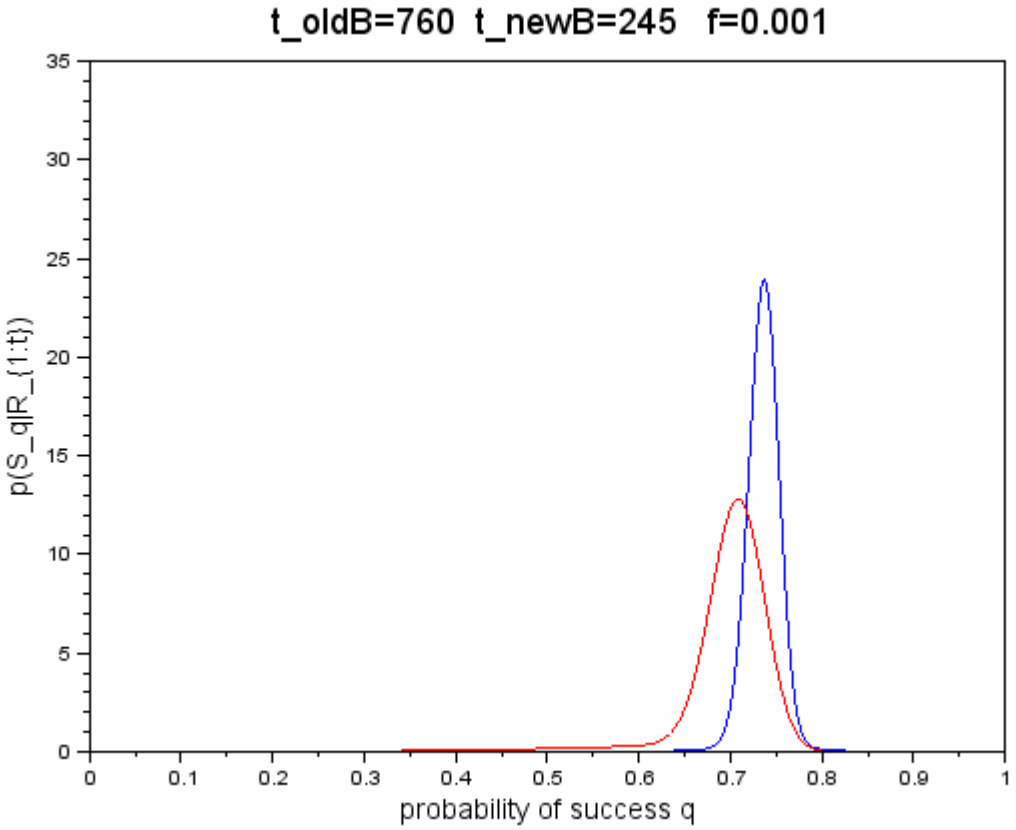
$$p(S_q|R_{1:t+1}I) \propto p(S_q|R_{1:t}I)\, p(R_{t+1}|S_q I) =$$
$$= [\, p(S_q E|R_{1:t}I) + p(S_q \bar{E}|R_{1:t}I)\,]\, p(R_{t+1}|S_q I) =$$
$$[\, \underbrace{p(E|R_{1:t}I)}_{f}\, \underbrace{p(S_q|R_{1:t}E\,I)}_{p(S_q|I)} + \underbrace{p(\bar{E}|R_{1:t}I)}_{(1-f)}\, p(S_q|R_{1:t}\bar{E}\,I)\,]\, p(R_{t+1}|S_q I)$$

Recursive formula:

$$p(S_q|R_{1:t+1}I) \propto [\, f\; p(S_q|I) + (1-f)\, p(S_q|R_{1:t}I)\,]\, p(R_{t+1}|S_q I)$$

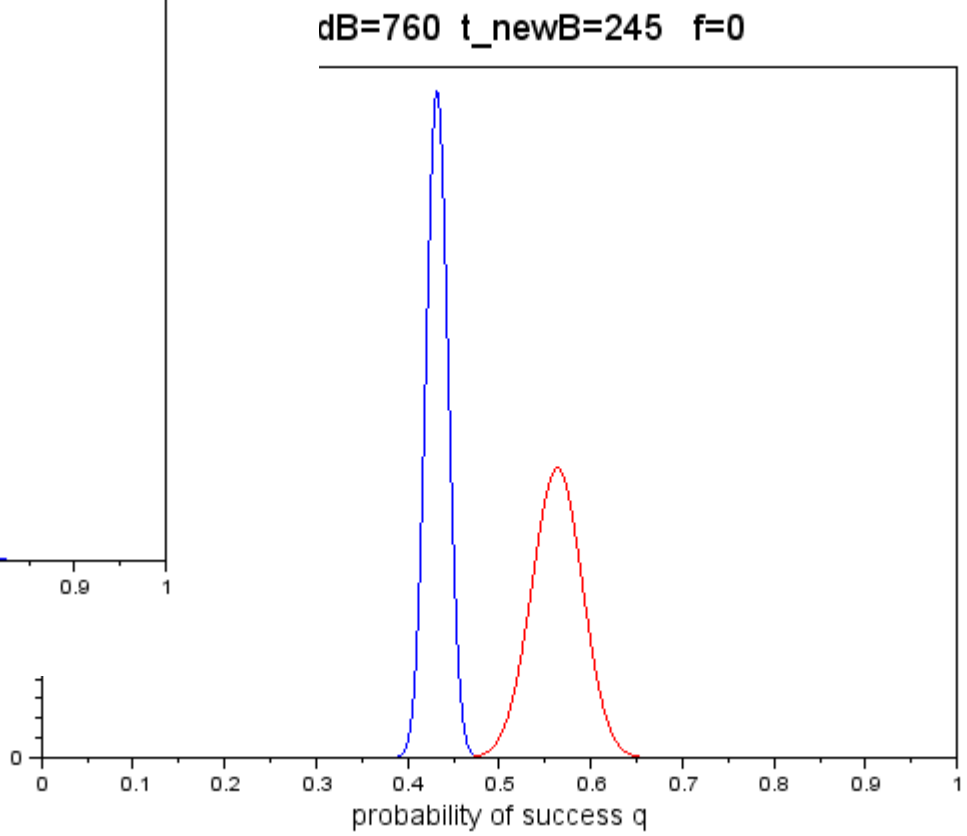
For f=0 we get the previous binomial distribution, but even for low values of f...

A probability distributions "race": the winner is...



Old method in blu

New method in red



Two completely different final results and sensitivities to data!

From a simple example to a more complex application...

Some complications for the cooperative data:

1. No binary results, all the instrument responses are possible

2. The extra factor to be considered is the possible instrument malfunctioning

3. The prior to be used for the observations is given by the forecast values, never uniformly distributed

4. The computational effort to manage probability distributions instead of single values is severe (in some cases prohibitive)

5. In operational conditions the available time for all the elaboration is short (a few hours)

Anyway the rules of Probability Theory and Inference remain the right tools to manage information in a logically consistent manner

Uncertainty estimation for cooperative data

$\tilde{y}$ = "the measure result for the measurand Y is $\tilde{y}$"

y = "within a dy y is the actual value of the measurand Y"

C = "the instrument measuring Y works correctly" (functioning)

$\bar{C}$ = "the instrument measuring Y does not work correctly" (malfunctioning)

Goal: find a computable expression for the probability density $p(y|\tilde{y}A)$, where A denotes all the available auxiliary information (instrument features, measure time and location, climatology, etc.)

$$p(y|\tilde{y}\,A) = p(yC|\tilde{y}\,A) + p(y\,\bar{C}|\tilde{y}\,A)$$

by product rule and

Hp1: $p(y|CA) = p(y|A) = p(y|\bar{C}\,A)$

Hp2: $p(\tilde{y}|y\,\bar{C}\,A) = p(\tilde{y}|\bar{C}\,A)$

Probability of functioning instrument

instrument response distribution (e.g. known by calibration or specification)

malfunctioning instrument response distribution

$$p(y|\tilde{y}\,A) \propto \left[\frac{P(C|A)}{1-P(C|A)} \frac{p(\tilde{y}|y\,C\,A)}{p(\tilde{y}|\bar{C}\,A)} + 1 \right] p(y|A)$$

prior for y (e.g. known by forecast)

Uncertainty estimation for cooperative data

$$p(y|\tilde{y}\,A) \propto \left[\frac{P(C|A)}{1-P(C|A)} \frac{p(\tilde{y}|y\,C\,A)}{p(\tilde{y}|\bar{C}\,A)} + 1 \right] p(y|A)$$

Some nice (and useful) properties of this formula:

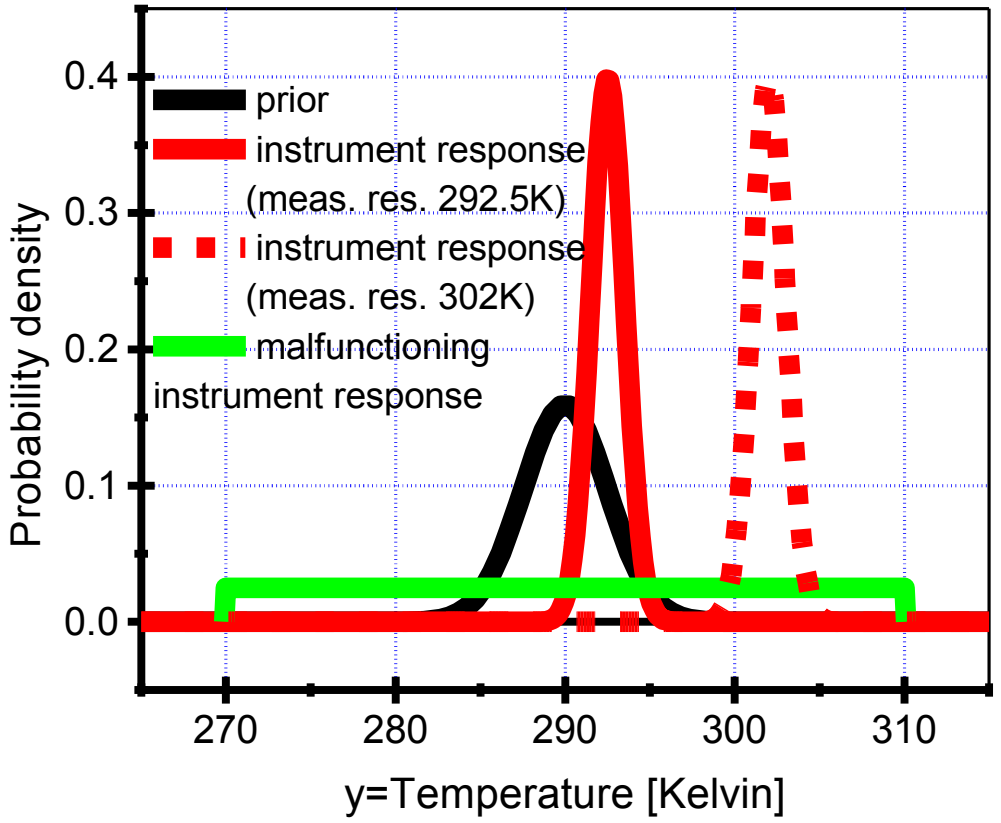
☺ It automatically collapses to the prior $p(y|A)$ when the instrument is almost surely broken, i.e. $P(C|A) \rightarrow 0$, and to the usual posterior probability when it surely works correctly, i.e. $P(C|A) \rightarrow 1$

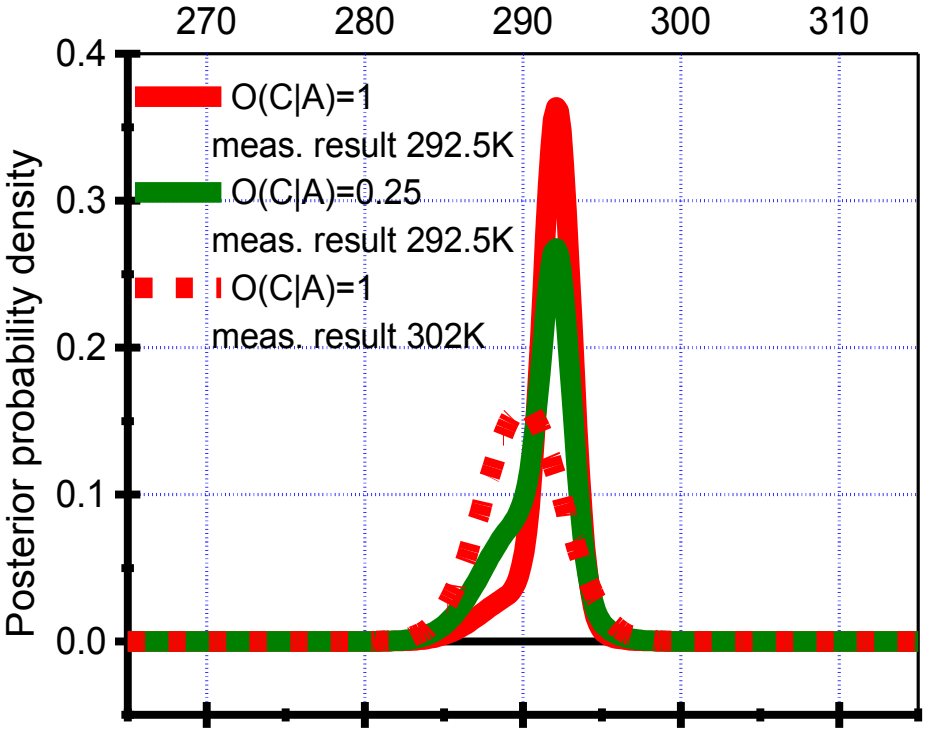
☺ It weights properly any intermediate case, allowing the exploitation of each single result and avoiding the necessity of rejecting it when the probability of malfunctioning exceeds some arbitrary threshold

☺ For any $P(C|A)$ it automatically tends to the prior when the measure result is largely implausible

☺ It easily accounts for possible malfunctioning warning flags (e.g. $\tilde{y}$ = -9999), simply assigning zero value to $p(\tilde{y}=\text{flag}|yCA)$ (or very a large value to the malfunctioning instrument response)

Uncertainty estimation for cooperative data





Conclusion

"To make a mickle from many a little, save your money and don't waste the collected data!"

Thank you for your attention