Analytical tools to deal with real systems

Laurent Dubroca

Ifremer

Musing on the concept of Good Environmental Status: the complexity of the status & the status of complexity

Joint Workshop 01/12/2020

Background

The Marine Strategy Directive Framework in a nutshell

- ob˙ective : good environmental status of the European marine waters = management of human activities while enabling a sustainable use of marine goods and services
- ecosystem management based on **indicators** to represent and track ecosystem status
- 11 descriptors (biodiversity, food-web, contaminants, eutrophication…)
- indicators (or combination of indicators) define a good environmental status

The Marine Strategy Directive Framework in practice

- an **indicator** : a data stream representing the process of interest
- an environmental status: **summarize** the information variability using a numerical treatment (a metric) into one single categorical value (good, bad etc)

Background

The Marine Strategy Directive Framework in practice : an example

- the eutrophication descriptor 5:
  - "Human-induced eutrophication is minimized [...] such as losses in biodiversity, ecosystem degradation, harmful algal blooms, and oxygen deficiency in bottom waters"
  - 8 criteria

- the criteria 2 (D5C2):
  - indicator: series of chlorophyll-a concentration
  - metric: the 90th percentile of the indicator
  - status: compare the metric to a threshold

Simple !

Question and warnings

As simple as the environmental status D5C2 may seem, wouldn't the devil hide in the details ?

In the *analytical application* of the details ?

Can operational complexity arise from the practical computation of one environmental status ?

Some warnings:

- The definition of the good environmental status given by D5C2 will be not questioned here.
- Different D5C2 computation settings coming from different papers are used to highlight the complexity of a practical computation of a simple indicator: these settings remains scientifically based and are not questioned !
- Some implementation details will not be take into account (national implementation...).

Complexity in practice

Good environmental status evaluation of D5C2 (English Channel for France highlighted, one reference):

- indicator: data of chlorophyll-a concentration based on satellite observation, aggregated in time and space
- metric: 90th percentile
- threshold: 1.44 mg.L-1

Simple

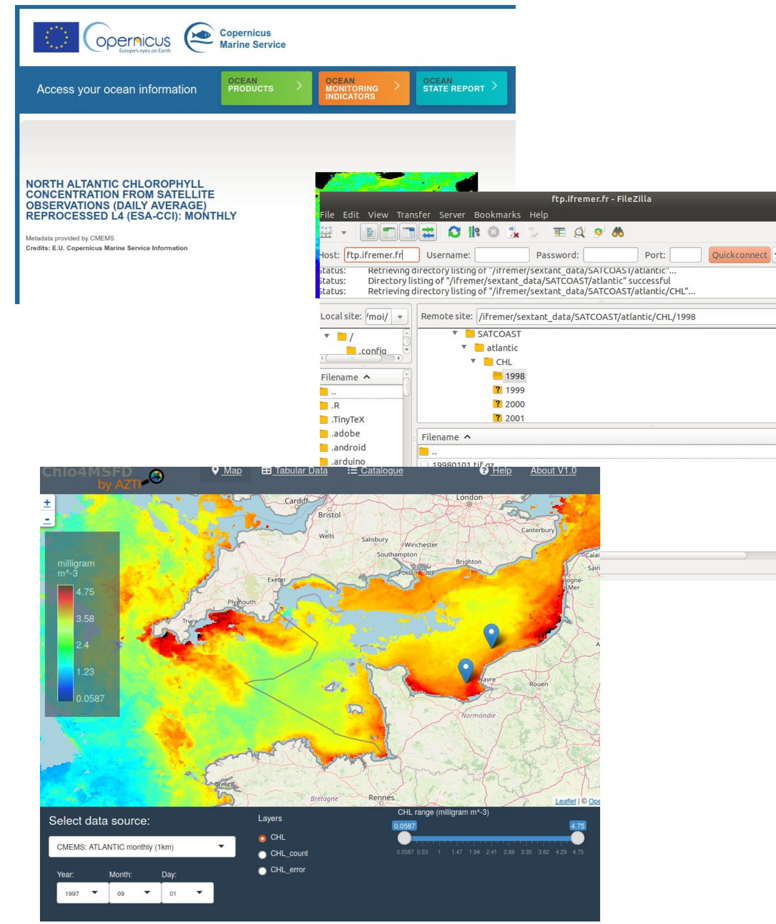
Criteria	D5C2 Chlorophyll-*a* in the Water Column
Indicators	[Chlorophyll-*a*]
Metrics	Percentile 90 of [Chlorophyll-*a*]
Elements Used (Parameters)	Surface chlorophyll-*a*
Metric Units	$\mu g \cdot L^{-1}$
RMU	All
Marine Regions	All
Assessment Years	CA: 2010–2015 IA and OA: 2010–2016
Data Aquisition Frequency	CA: twice a month IA and OA: daily
Assessment Periods	Atl.: March to October WMS: every months
Data Sources	CA: REPHY IA and OA: remote sensing (MODIS-OC5Me)
Assessment Thresholds	Atl.: CA NEA 1/26a: 10 µg·L-1 CA NEA 1/26b: 15 µg·L-1 IA NEA 1/26a: 5 µg·L-1 IA NEA 1/26b: 7.5 $\mu g \cdot L^{-1}$ OA NEA 1/26a: 4 $\mu g \cdot L^{-1}$ OA NEA 1/26b: 6 $\mu g \cdot L^{-1}$ WMS: CA Ecotype 1: 10 $\mu g \cdot L^{-1}$ CA Ecotype 2A: 3.6 $\mu g \cdot L^{-1}$ CA Ecotype 3W: 1.8 $\mu g \cdot L^{-1}$ CA Ecotype W: 1.22 $\mu g \cdot L^{-1}$ IA: 2 $\mu g \cdot L^{-1}$ OA: 1.44 $\mu g \cdot L^{-1}$

(from Lefebvre et al. 2020)

Complexity in practice : data access

Good environmental status evaluation of D5C2 adding other references (e.g. Bor˙a et al. 2019, Gohin et al. 2019):

- data:
    - Marine Copernicus data portal ?
    - Gohin et al. 2019 ftp ?
    - chl4MSFD ?
    - ...
- data extraction:
    - spatial resolution : 1km, 4km ?
    - daily, weekly, monthly values ?
    - interpolated data?
    - temporal range: 2009-2014 ? 2010-2016 ?



~~Simple~~

Complexity in practice : analyses

Good environmental status evaluation of D5C2 adding other references (e.g. Bor˙a et al. 2019, Gohin et al. 2019):

- for spatial or temporal aggregation: how to summarize the data ? How to compute uncertainties ?
- metric: quantile ?
  - there are 9 sample quantile methods available
  - what is the method to use ?
- thresholds for D5C2:
  - 1.44 $\mu g.L^{-1}$ (Lefebvre et al. 2020)
  - 1.52 $\mu g.L^{-1}$ (Bor˙a et al. 2019)
  - …

~~Simple~~

All sample quantiles are defined as weighted averages of consecutive order statistics. Sample quantiles of type *i* are defined by:

$$Q[i](p) = (1 - \gamma)\, x[j] + \gamma\, x[j+1],$$

where $1 \le i \le 9$, $(j-m)/n \le p < (j-m+1)/n$, $x[j]$ is the *j*th order statistic, *n* is the sample size, the value of γ is a function of $j = floor(np + m)$ and $g = np + m - j$, and *m* is a constant determined by the sample quantile type.

Discontinuous sample quantile types 1, 2, and 3

For types 1, 2 and 3, $Q[i](p)$ is a discontinuous function of *p*, with $m = 0$ when $i = 1$ and $i = 2$, and $m = -1/2$ when $i = 3$.

Type 1

Inverse of empirical distribution function. $\gamma = 0$ if $g = 0$, and 1 otherwise.

Type 2

Similar to type 1 but with averaging at discontinuities. $\gamma = 0.5$ if $g = 0$, and 1 otherwise.

Type 3

SAS definition: nearest even order statistic. $\gamma = 0$ if $g = 0$ and *j* is even, and 1 otherwise.

Continuous sample quantile types 4 through 9

For types 4 through 9, $Q[i](p)$ is a continuous function of *p*, with $gamma = g$ and *m* given below. The sample quantiles can be obtained equivalently by linear interpolation between the points $(p[k],x[k])$ where $x[k]$ is the *k*th order statistic. Specific expressions for $p[k]$ are given below.

Type 4

$m = 0$. $p[k] = k / n$. That is, linear interpolation of the empirical cdf.

Type 5

$m = 1/2$. $p[k] = (k - 0.5) / n$. That is a piecewise linear function where the knots are the values midway through the steps of the empirical cdf. This is popular amongst hydrologists.

Type 6

$m = p$. $p[k] = k / (n + 1)$. Thus $p[k] = E[F(x[k])]$. This is used by Minitab and by SPSS.

Type 7

$m = 1-p$. $p[k] = (k - 1) / (n - 1)$. In this case, $p[k] = mode[F(x[k])]$. This is used by S.

Type 8

$m = (p+1)/3$. $p[k] = (k - 1/3) / (n + 1/3)$. Then $p[k] =\sim median[F(x[k])]$. The resulting quantile estimates are approximately median-unbiased regardless of the distribution of `x`.

Type 9

$m = p/4 + 3/8$. $p[k] = (k - 3/8) / (n + 1/4)$. The resulting quantile estimates are approximately unbiased for the expected order statistics if `x` is normally distributed.

Further details are provided in Hyndman and Fan (1996) who recommended type 8. The default method is type 7, as used by S and by `R` < 2.0.0.

R programming language quantile help page

Complexity in practice : a summary

- data:
  - definition unclear
  - tedious access
- analyses:
  - data aggregation
  - metric
  - status computation

~~Simple~~
Complexity, at last.

- In short
  - data definition and access
  - replicability

Complexity in practice : solutions ?

Marine data access

- open data (use of European marine data portal or publish them)
- identify clearly the data source (DOI)

- access the data using web service

25.4.2007 EN Official Journal of the European Union L 108/1

DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 14 March 2007

establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)

Complexity in practice : solutions ?

Replicability:

- Analytical methods to assess good environmental status should be transparent
- publish the code !
  - an example : the NEAT (Nested Environmental status Assessment Tool) software developed during the DEVOTES pro˙ect

Analytical tools:

- open source programming language (R, python…)
- dynamic report generation : text + code (Jupyter notebook, Rmarkdown…)
- use content version system (github, gitlab) to track and share the change in the analytical processes

Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

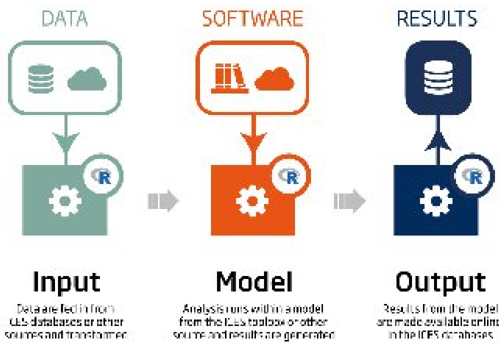
	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

Transparency and openess promotion guidelines from Nosek et al. 2015

Complexity in practice : example from the fishery science world

- The International Council for the Exploration of the Sea (ICES) provides advice for management of wild capture fisheries in the European waters.
- these advices feed the MSFD descriptor D3 (commercial exploited fish and shellfish) providing two indicators: the fishing pressure and the stock biomass data.

- **open data and web services**
    - the ICES stock information and stock assessment databases provide web services to access the data and stock assessment data, with support to the R programming language.
- **replicability**
    - fishing pressure and stock biomass are computed during stock assessments. The Transparent Assessment Framework (TAF) is an online open resource of ICES stock assessments for each assessment year. All data input and output is fully traceable and versioned.



Conclusion: analytical tools to deal with real systems ?

- Open data, web services and replicability are the keys to avoid complexity in single good environmental status computation.

- The analytical tools to deal with this complexity:
  - European marine portals and open data
  - web services to interface data and analyses
  - open source tools (programming language and content version system) to share and review the good environmental status computation

~~Simple~~ → ~~Complex~~ → Manageable reality

JPI OCEANS

THANK YOU

Laurent Dubroca

Bibliography

- Bor˙a, Angel, Joxe Mikel Garmendia, Iratxe Menchaca, Ainhize Uriarte, and Yolanda Sagarmínaga. 2019. “Yes, We Can! Large-Scale Integrative Assessment of European Regional Seas, Using Open Access Databases.” Frontiers in Marine Science 6. https://doi.org/10.3389/fmars.2019.00019.
- Gohin, Francis, Dimitry Van der Zande, Gavin Tilstone, Marieke A. Eleveld, Alain Lefebvre, Francoise Andrieux-Loyer, Anouk N. Blauw, et al. 2019. “Twenty Years of Satellite and in Situ Observations of Surface Chlorophyll-a from the Northern Bay of Biscay to the Eastern English Channel. Is the Water Quality Improving?” Remote Sensing of Environment 233 (November): 111343. https://doi.org/10.1016/˙.rse.2019.111343.
- Lefebvre, Alain, and David Devreker. 2020. “First Comprehensive Quantitative Multi-Parameter Assessment of the Eutrophication Status from Coastal to Marine French Waters in the English Channel, the Celtic Sea, the Bay of Biscay, and the Mediterranean Sea.” Journal of Marine Science and Engineering 8 (8): 561. https://doi.org/10.3390/˙mse8080561.
- Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. “Promoting an Open Research Culture.” Science 348 (6242): 1422–25. https://doi.org/10.1126/science.aab2374.

- Quantiles methods:
  - Hyndman, Rob, and Yanan Fan. 1996. “Sample Quantiles in Statistical Packages.” The American Statistician 50 (November): 361–65. https://doi.org/10.1080/00031305.1996.10473566.
  - https://en.wikipedia.org/wiki/Quantile, accessed the 7/11/2020.