Comments on the OHI Documentation and Tools for a Regional Study

Soraya Abad-Mota
With the contributions of: Lelys Bravo, Mario Hurtado,
Geovagi Flores and María Belén Segovia

August 2015

1. Introduction

The consortium Consulsua/Biotica was formed to develop a regional study to measure the ocean health index in the Gulf of Guayaquil, Ecuador, starting at the end of 2014. The goal of this study is to apply the OHI methodolgy developed by Ben Halpern in U.C. Santa Barbara, to measure the ocean health in that region of Ecuador.

Within the consortium a technical group of Information Management (IM) was formed. The IM group started with three members and grew to six members at the final stages of the project. The first task of the group was to understand the steps of the methodology and provide support to the field consultants in following the methodology. The field consultants were specialists in different areas, which include biologists, oceanographers, economists, sociologists, among others. There were approximately 14 field consultants in total, participating at different stages of the project. There were two coordinators, one for the administrative aspects and another for the technical and specialized tasks.

The IM group received, classified and organized the data collected by the field consultants. These data are used to create the appropriate data layers necessary for computing the scores of each goal and the global ocean index of the Golfo de Guayaquil region. In addition to the data management tasks, the IM group analyzed the github structure created for the country, developed a new structure for this regional study and read the existing R programs to manipulate the data layers and calculate the scores. To update the calculations of scores, trends and status, and to manage the data, several new R routines were written and included in the Git platform.

This document presents some comments and conclusions of members of the IM team and the Project's technical coordinator, about the use of the OHI framework and its tools in the Gulf of Guayaquil, and provides some recommendations for future regional study groups.

2. Documentation

The main documents used in this project are:

- Supplementary Information to the paper: Halpern, B. S., Longo, C., Hardy, D., McLeod, K. L., Samhouri, J. F., Katona, S. K., ... & Zeller, D. (2012). An index to assess the health and benefits of the global ocean. Nature, 488(7413), 615-620. The group usually referred to this document as Halpern 2012.

- The OHI manual. At first a static version of this document was used but since it is always being updated for improvements, the online version was used instead, it is available at ohi-science.org/manual.

To a less extent the following documents were used:

- Guía Conceptual del índice de la salud del océano: Filosofía y Marco. Versión 2 - Abril 2014. Mostly at the beginning, when learning about the methodology.

- Halpern BS, Longo C, Scarborough C, Hardy D, Best BD, Doney SC, et al. (2014) Assessing the Health of the U.S. West Coast with a Regional-Scale Application of the Ocean Health Index. PLoS ONE 9(6): e98995. doi:10.1371/journal.pone.0098995 To clarify several aspects of a regional study.

- Elfes, C. T., Longo, C., Halpern, B. S., Hardy, D., Scarborough, C., Best, B. D., Tiago Pinheiro & Dutra, G. F. (2014). A regional-scale ocean health index for Brazil. PloS one, 9(4), e92589. Useful at the end, when writing the final reports. It was a guide of what is included and highlighted in a regional study.

The documentation available was somewhat overwhelming, but it was obvious that Halpern 2012 is a very important and useful source all along the project because it describes the whole OHI methodology. In spite of this, in many ocassions the explanations on how to calculate particular data layers were obscure. This should be improved.

The study of each data layer is extremely important, specially if local data is to replace the data used in the global assessments. It is necessary not only to understand the philosophy of each goal but the data layers associated to each goal.

Recommendations:

1. Distinguish between the actual methodology and the advice on how to apply it. The group should know in which aspects there is freedom to change without violating the rules and standards of the methodology.

2. Emphasize the existing relationships between data layers, this is not highlighted enough in the documentation. For example, the value of the resilience data layer

wgi_all is computed by evaluating the expression $1 - ss\_wgi$, in other words, the sum of both data layers is always 1: $ss\_wgi + wgi\_all = 1$. The same happens with some pressure data layers and the corresponding layer used for the status of a goal, for example, *po_pathogens* is used to determine both, the status of the Clean Waters goal and a pressure, if *po_pathogens* is computed as a pressure, then $1 - po\_pathogens$ is the value for the status. All the relationships between data layers should be made more explicit in the documentation.

3. In some cases it was hard to find complete data to compute the trends for the last five consecutive years. It could be explored if it is possible to relax some of the constraints on computing the trends.

4. More specific comments about the data layers follow:

   a) There is a variety of calculations necessary to compute the values of the data layers, some require a complex mathematical procedure, for example *hd_subtidal_sb*, and for others the computation is fairly straightforward, as is the case of *hd_intertidal*. Many field consultants did not have a strong mathematical background and for them it was hard to understand the complex procedures. Some of these complex computations are extensively explained in the documentation, others are not and also there are some cases where the external sources for the explanation were no longer available. For an example see the description of *fp_com_hb*.

   b) Some data layers which measure different aspects, use the same base information and sources, for example, *ss_wgi* and *wgi_all* both use the Worldwide Governance Indicators (WGI), therefore it would be useful to handle them simultaneously to avoid effort duplication.

   c) It is not clear when it is possible to leave blanks or null values in a data layer or when they should be replaced by zeroes, for the data layers where this could happen. The difference between a null value and a zero is conceptually important but for some data layers the null values are not allowed and they must be replaced by zeroes, as is the case for *fis_meancatch*. These situations are not documented in the OHI material we used.

   d) When the value of a data layer is a score it is usually not docummented how to calculate that score.

   e) The pressure data layers are particularly complex, their values are scores in the range 0,1 but each of them has a reference point, it is not always clear which reference point to use when computing the pressure value.

5. It would be good to include a checklist of what is important to report in a regional study, for example, the percentage of data layers filled with local data or the percentage of goals for which the model was changed.

6. It is also useful to provide a list of caveats to take into account when reporting about a regional study, for instance, it is not fair to compare the scores of a global study which uses global data with the scores of a regional study.

7. Pressures and resiliencies deserve specific comments when finding data to calculate their values in a regional study. It was hard some times to find data at the local or provincial level. Most regulations on the environment, pollution and fisheries are global for the whole country, therefore it is frequently impossible for the local authorities to provide solutions or to find disaggregated data. To evaluate resiliencies, the OHI methodology suggests the use of the questionnaires provided by CBD; for Ecuador the answers to the questions in those surveys are from 2005 and rather obsolete. For this study we had to use different sources and be very creative about it in order to find appropriate indicators.

3. The Tools

The recommendations about the tools available in the OHI methodology are grouped into three categories: the toolbox, the application and the R code provided by the framework.

Toolbox

It would be useful to have a top-down vision or diagram of all the tools and their use in the framework. This has already been included in the most recent version of the OHI manual (Aug. 24, 2015). This diagram shows which are the tools included in the *Toolbox*: the **R programming language**, **R-studio** as a preferred interface to R, **Git** as the version control software to maintain the data and the R code and **GitHub** which is the interface to Git.

The use of the toolbox software starts with the preparation of the data layers, but before, in the process of discovering and gathering input information, there are no tools or suggestions of which tools to use, also there is no advice on how to organize the data. Even though this stage should be determined by each group, it would be useful to make some general suggestions in the manual, or at least acknowledge the fact that this is a phase were the group must spend some time planning and finding appropriate means of storing and documenting the acquired data.

In the Guayaquil IM team we developed a process which starts with the acquisition of the raw data and ends with the inclusion of the data in git in the prep folder. We also differentiatie the tasks of grouping the data in the folder for each goal from the final production of the data layers that feed the score and index calculations, each of these tasks has a separate folder structure in Git. It is important to notice that the organization of the data layers by goal is a little confusing and prone to errors, since many data layers are shared by different goals and some pressures are also used in status calculations. It is preferrable to organize the data, by groups of data layers. Maybe it would also be more convenient to separate into two layers, those that are used as a pressure and as a status indicator.

During the Guayaquil study we first thought of using *Github* earlier, but since the data gathered was mostly in excel and pdf files, it was inconvenient to use Github for those formats. We decided to use dropbox to store and share those files. The important feature of a tool for this early stage is an easy way to organize the incoming data and reliability on the data stored, through a mechanism of backups or a reliable storage system.

Application ohi-science/gye/app

The application to visualize the data layers and scores is very useful, all along the project until the end, it was consulted and used often. We suggest to emphasize more that this application is mainly a visualization tool for the data, metadata and scores. It is very useful, but the non-computing specialists get extremely confused between the Web App and the github platform.

On the other hand, the gye app solely was not helpful to understand the calculations, for this it is necessary to understand the R code contained in functions.R

It is also important to add the following:

1. Expand the descriptions of the data layers. We found examples of data layers for which the units were not clear or had imprecisions and it was time consuming to clear those out. The contents of the application documents could be cleaned and revised to improve these descriptions.
2. Instructions on how to update the documentation contained in the application: description of the model and description of new data layers. Include this step as part of the tasks performed by the group.
3. How to answer database questions on the data layers and the goals. During the Guayaquil study there were several ocassions in which the IM group was asked about this kind of questions, some examples of these questions follow. Which data layers are used for status only? which are pressures? which data layers are used for both status and pressure? Which data layers are common to several goals and how many are there? Which data layers are used to determine the status of a specific goal? How many data layers are used in a specific goal? Which kind of reference point is used for this goal and how is it used in the calculations?

The R code

We noticed some aspects of the R code included in the OHI assessment repository that we want to highlight.

1. Make more explicit the calculations of reference points, current status, most likely future and trends. It is rather obscure for the users where the reference points are calculated.
2. The treatment of reference points in the code could be made more uniform.
3. Explain the difference between the OHI core routines and the OHI Assessment routines and list which calculations belong to each.

4. In general, it is noticeable that the code was written by different groups. It is advisable to make the coding more uniform. This will make it more readable and will increase maintainability.

4. Recommendations on how to use the framework

1. Emphasize the importance of considering all the data layers at once from the beginning of the study. Do not leave the pressures and resiliences for the end.
2. In some sense, when collecting data the distinction between goals is not that important, rather the determination of which data layers are relevant for the study is crucial. Many data layers are shared by several goals.
3. In our experience, it is better to work and build the data layers by *blocks*, assembling ALL related data layers or all the layers that belong to the same goal, simultaneously.
4. There is always an interplay between pressures and resiliences, no data layer in those contexts should be studied in isolation.

5. Practical Extensions in the use of the OHI framework

In our experience of applying the OHI methodology in the regional study of the Gulf of Guayaquil, we established a few extensions to the main methodology that might be useful for other groups. What follows is a list of such extensions.

- We extended the content and use of the README files for the data layers, they constitute important documentation sources and could be made as rich in content as needed. We included the following nine items in the README files:
    1. Data collection Institution
    2. Institution providing the data to this project (if different from the original)
    3. Cites to the source of the data (scientific papers)
    4. Data capture methods, i.e. details of the study, instrument, method, survey or form used.
    5. Period of time covered by the data
    6. Unit of the values
    7. Meaning of the values and how to interpret them.
    8. Relevant comments about the data or procedure to compute the values.

9. Meaning of the columns, if the data layer is new, created for this study or if different from the description contained in the OHI-ecu application.

It is also important to explain details of the calculations to produce the values of each data layer. This could be included in the relevant comments on item 8 above or create a new section for those details.

- We established more standards for file names and writing the reports and the readme files. In general, the OHI manual could encourage the members of the technical group to establish their own standards, early in the project. In the github/gye repository we have included a file called A-instructivoCapasDatos.pdf which describes some of these standards used in the Gulf study.

- We found that the cloning of a repository was a task in which the members of the team needed help, so we developed a detailed example of how to do it and included it in the github/gye repository, in the file A-ClonarRepo.pdf.

- An observation we received frequently was about the database model used by the framework, but since it was not designed using the formal notion of a database, there was not a conceptual database model associated with the methodology, except for a couple of old diagrams obtained from: ftp://ohi.nceas.ucsb.edu/pub/data/2012/. On a next phase of the OHI project it might be useful to consider the inclusion of a database view of the data layers.

After reflecting on the process of evaluating the OHI in the Gulf of Guayaquil we think it is important to have a second phase for the implementation of policies, folllow-up and evaluation of the ocean health index estimated in the first phase.