Employing software engineering principles to enhance analysis of coral reef databases

Mark Jenne, M.M. Dalkilic, Claudia C. Johnson

Abstract The challenges presented by data to scientific inquiry and hypothesis testing in an oceanographic setting are not new problems. Indeed, the challenges are at least a century old. The problems are not with the data itself, but rather with a lack of attention to the management of the "data ecology" in the information systems. Data needs to be accessible as an input to scientific inquiry – a requirement that goes far beyond simply centralizing the available data. Our research focuses on the development of a proof-of-concept system that properly handles an information ecology. The power of such a foundation is then demonstrated through our data driven hypothesis generation system, built and employed for analysis of the relationship between coral disease and temperature in the Caribbean.

Keywords: big data, software engineering, database, data-driven, coral disease

---

Mark Jenne, mjenne@indiana.edu
Computer Science, Indiana University, Bloomington, IN, USA

M.M. Dalkilic, dalkilic@indiana.edu
Computer Science, Indiana University, Bloomington, IN, USA

Claudia C. Johnson, claudia@indiana.edu
Department of Geological Sciences, Indiana University, Bloomington, IN, USA

Corresponding author: Mark Jenne, mjenne@indiana.edu

Introduction

Many oceanographic data repositories have come online in the last few decades. Some repositories are large oceanographic datasets (World Ocean Database (WOD) (Boyer 2013) and World Ocean Atlas (WOA) (Locarnini 2013)), while others have more specific content (ReefBase Coral Bleaching GIS (WorldFish 2016) and Global Coral Disease Database (GCDD 2016)). Data stored in these repositories, particularly the WOD and WOA, are vast and invaluable. Making data accessible as a product itself, however, is not sufficient for driving

large-scale analyses needed to affect policy decisions. The focus of our research is to make data available for scientific inquiry in a data-driven approach. We recognize that significant effort and resources are required to ensure longevity of data in an information system. Our proof-of-concept software for data governance and robust management of the data ecology in our information system is developed with this in mind. Furthermore, we demonstrate the power of such an approach by building an algorithm for a high-level analysis of coral disease in relation to ocean temperature in the Caribbean on top of our information ecology framework that uses not thousands, but millions of data points. Together, the components of our system allow for programmatic search of the coral disease and temperature space to form testable hypotheses – a methodology referred to as data-driven hypothesis generation. It is this data-driven approach that allows us to perform large-scale analyses needed to address the questions looming large for coral reefs, and to bridge the gap between science and policy.

Materials and methods

The software systems behind the two primary components in this research are: (1) the information ecology framework and (2) an algorithm for the preliminary analysis of coral disease and temperature. We include only an overview of these components here. The configuration or experimental parameterization for the system behind the data-driven hypothesis generation is then provided.

Rather than following the traditional computational science approach of ushering all of the data to the algorithm and then building out the algorithm to incorporate all elements of the transformation, management, and processing of the data, we isolate the non-research related procedures and push them to the data. Together, this set of procedures and the controlling software around them, form the backbone of what we are calling our information "ecology framework." This system accomplishes two major goals: (1) provisioning a robust data manager with all necessary extraction, transformation, and loading components and (2) isolating the processes involved in scientific inquiry algorithm development.

With the data ecology framework in place, the algorithmic component focuses on the search of the data space and extraction of isolated relationships between the data for hypothesis testing (Fig. 1). As a first pass, our approach leverages a naïve quasi-clustering technique for

establishing spatial bounds for the geographic extent of coral disease outbreaks and is followed by a search of the large data space of sea temperature for local representative temperature data.

The newly associated data are then used to produce visual analytic tools for expert analysis from which individual, testable hypotheses are extracted for further consideration (Fig. 2). This methodology employs data-driven hypothesis generation by trading in the necessity for explicit, testable claims to drive experimental setup in favor of general pattern search within the data pertaining to the relationship between coral disease and temperature.

Our analysis of the relationship between coral disease and temperature was constrained to the Caribbean during the years 1970 to 2009. The geographic locations of interest in this study were produced through clustering of the individual coral disease records using a spatial threshold of 10 km. Temperature is represented by average annual near-surface sea temperature regional to the disease clusters, via an expanding local search with a contingency fallback to a Caribbean-wide annual average. Missing data in the temperature catalogue is thus dealt with at the same regional scale and represents another small, but significant advantage provided by the information ecology framework: effective policy implementation at the data-level rather than at the algorithmic-level. . In Fig. 2, temperature is parameterized at a depth of $0.0m$, but analyses including temperature data at additional depths are easily achieved in this context.

Results

The results presented here are in the context of both a big data problem and large-scale analyses. Making use of our data ecology framework, our algorithm for data-driven hypothesis generation regarding the temperature – coral disease relationship in the Caribbean was able to integrate and process the complete coral disease catalogue presented by ReefBase and the complete ocean temperature data set hosted in the WOD. Grouping the 5,038 coral disease records into spatial clusters yielded 293 distinct geographic locations for analysis. At each location, respective temperature data subsets were selected from the more than 62 million data points available.

Algorithm 1 Coral Disease Temperature Analysis

INPUT data $\{\Delta_1, \Delta_2\}$, config Φ
OUTPUT Temperature–Disease Timelines $A_1, \ldots, A_n \in \mathsf{A}$
%% assume that each A_i is a tuple $(lat, lon, \mathsf{D} \in \Delta_1, \mathsf{T} \in \Delta_2, \mathsf{Y})$
%% where each $Y_i \in \mathsf{Y}$ is $(y, D_i \subset \mathsf{D}, T_i \subset \mathsf{T}, \mathsf{C})$
%% y is the year, D_i is the subset of coral diseases at this location
%% T_i is the subset of temperatures at this location
%% C is the list of corals affected by disease at this location
arrange disease records temporally $\Delta_1' \leftarrow \Delta_1$
construct empty set A
%% cluster disease instances
for $\mathbf{x} \in \Delta_1'$ **do**
  flag $\leftarrow$ false
  for $A_i \in \mathsf{A}$ **do**
    if $\mathbf{x}$.distance$(A_i.lat, A_i.lon) \leqslant \Phi.radius$ **then**
      $A_i.\mathsf{D}.add(\mathbf{x})$
      $flag \leftarrow true$
    end if
  end for
  if !flag **then**
    %% add a new Temperature-Disease Timeline at the geological location of this disease
    $\mathsf{A} \leftarrow \mathsf{A} \cup A(\mathbf{x})$
  end if
end for
%% associate temperature data with disease clusters
for $A_i \in \mathsf{A}$ **do**
  for $D_j.y \in A_i.\mathsf{D}$ **do**
    $itr \leftarrow 0$
    while $\|resultSet\| = 0 \wedge itr < \Phi.maxItr$ **do**
      $resultSet \leftarrow SpatialQuery(\Delta_2, A_i.lat, A_i.lon, D_j.y, \Phi.radius + (\Phi.radius * (itr/2)))$
    end while
    if $\|resultSet\| = 0$ **then**
      $A_i.\mathsf{T} \leftarrow SpatialQuery(\Delta_2, D_j.y)$
    else
      $A_i.\mathsf{T} \leftarrow resultSet$
    end if
  end for
end for
%% process temperature-disease timelines
for $A_i \in \mathsf{A}$ **do**
  for $Y_j \in A_i.\mathsf{Y}$ **do**
    $Y_i.D_i \leftarrow \mathsf{D}.Where(x => x.year = Y_i.y)$
    $Y_i.T_i \leftarrow \mathsf{T}.Where(x => x.year = Y_i.y)$
    $Y_i.\mathsf{C} \leftarrow D_i.Select(x.genusSpecies => x)$
  end for
end for
$OutputViz(\mathsf{A})$

Fig. 1 Formal notation of the algorithm developed for the Caribbean-wide analysis of the relationship between temperature and coral disease

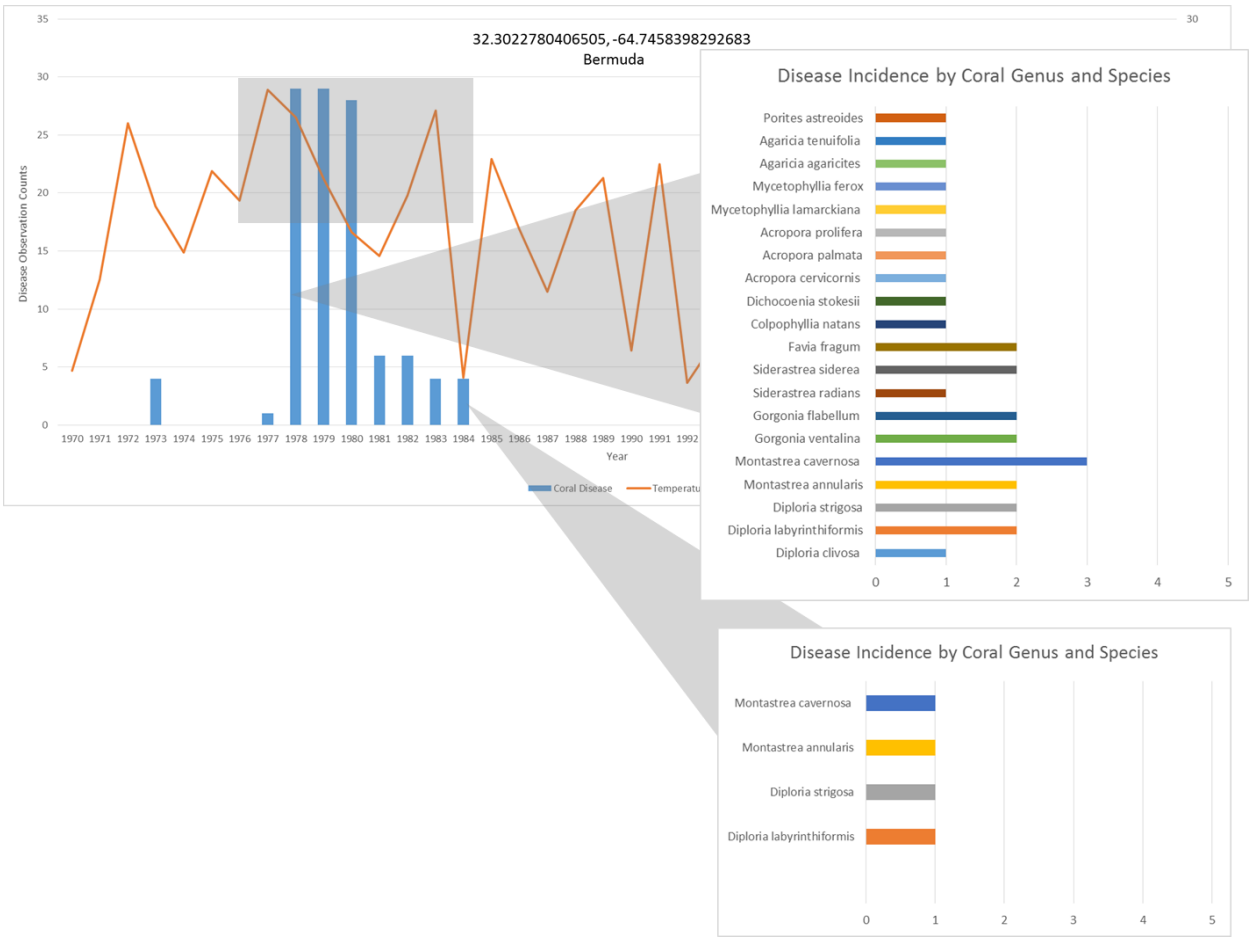

Fig. 2 Coral disease and temperature analysis performed at a geographic location near the Florida Keys. Seen here are the high-level trends, a sample distribution of diseased coral taken from 1991, and the isolated patterns of interest from which testable hypotheses regarding the relationship between temperaeture and coral disease are generated

The resulting coral disease and temperature sets were aligned temporally visualized for extraction of testable hypotheses.

Produced from a disease cluster near Key Largo (cluster center located at 25.1383, -80.3029), Fig. 2 provides an example of the rich information available in the software context of each individual analysis. Here we see the relationship between regional temperature and coral disease for this location from 1970 to 2009. An example of the kind of finer-grained information that is available is presented in the distribution of corals affected by disease in 1991 by genus and species. Finally, and most importantly, we have identified three regions of interest according to the data visualization. The high-lighted regions 1, 2, and 3 represent a potential set of data-driven hypotheses derived from the patterns in the data. Consistent with existing hypotheses regarding the relationship between temperature and coral disease in the literature, we have (1) thermal stress as high variance in temperature causes disease outbreak (McClanahan, 2007), (2) thermal stress as highly elevated temperature causes significant outbreak of coral disease (Jones, 2004), and (3) consistently elevated temperature favors disease persistence (Porter, 2001).

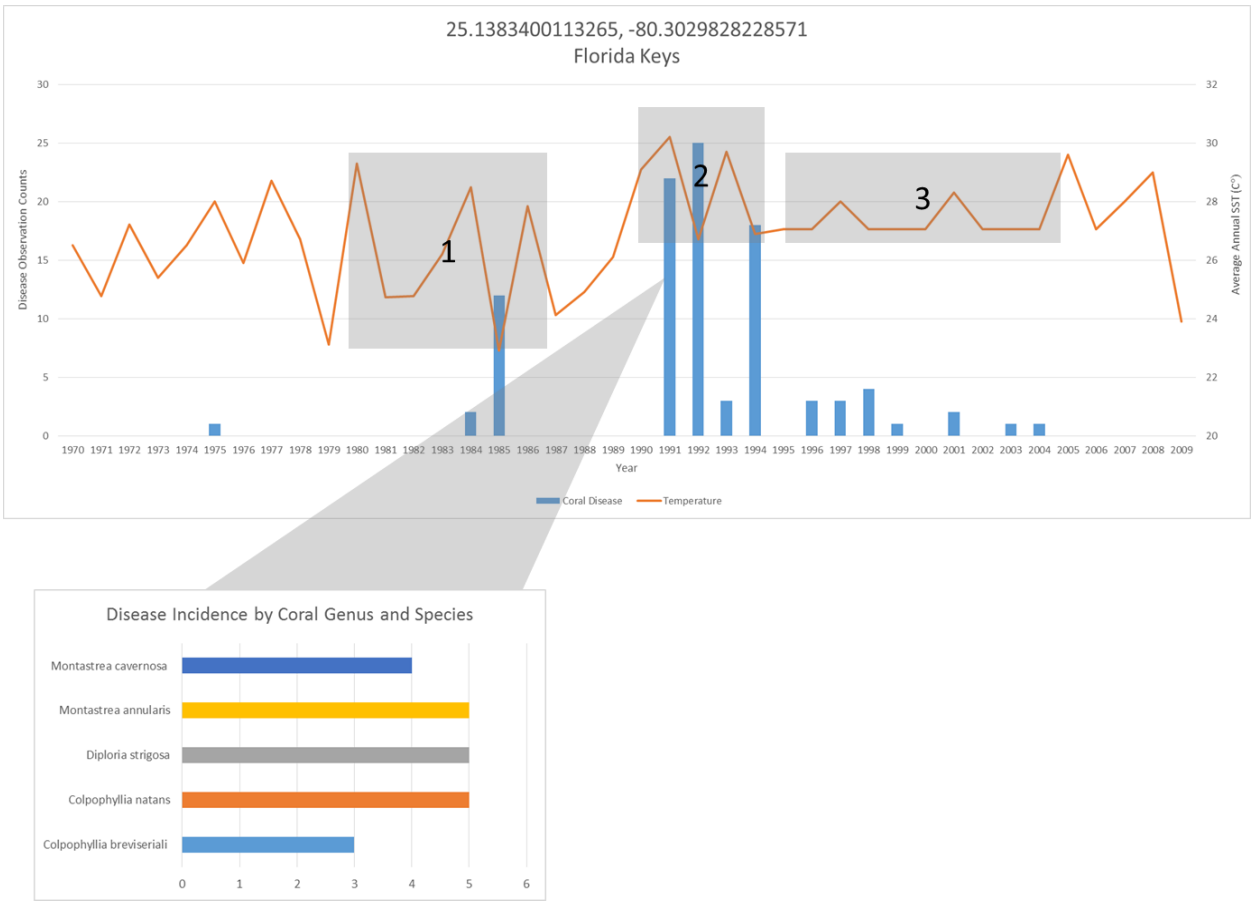

Fig. 3 Coral disease and temperature analysis performed at a geographic location near Bermuda. The pattern of interest here is the significant difference in corals affected by disesase following similar thermal stress in 1977 and 1983

Another geographic location for the temperature, coral disease analysis is shown in Fig. 3. For this location we again see the complete time series for regional temperature and coral disease for this location from 1970 to 2009. Here, there is an interesting pattern underlying the change in the distribution of corals affected by disease following two similar instances of thermal stress in 1977 and 1983. The dramatic shift in the corals affected by disease from 1978 to 1984 may provide isolated data that can begin to explore the possible effect of coral community adaptation to environmental stresses such as disease. Should the data reflect that this dramatic shift resulted from significant coral death, then an additional data-driven hypothesis could be that the remaining corals are more acclimated to the thermal stress and that when they co-occur they increase the overall resilience of the reef.

A final example of data-driven hypothesis generation using the combination of our information ecology framework and algorithm can be seen in Figure 4. Representing a

geographic location near Curaçoa, we have the complete disease and temperature time series standing alone without any finer-grained information. In this case, the data reveal what may be an anti-pattern in the data – a case that is not consistent with existing information or accepted hypotheses. From 1996 through 2008 there appears to be persistence of coral disease at this location despite little deviation from the periodicity of the temperature trend during the previous25 years. More subtle environmental factors may be involved – eutrophication or other effects – that are creating a state of continual disturbance for the reef allowing for the disease persistence. This instance further reveals the overall strength of our joint system, as we have been able to isolate a data set that may yield valuable hypotheses beyond the relationship between temperature and disease.

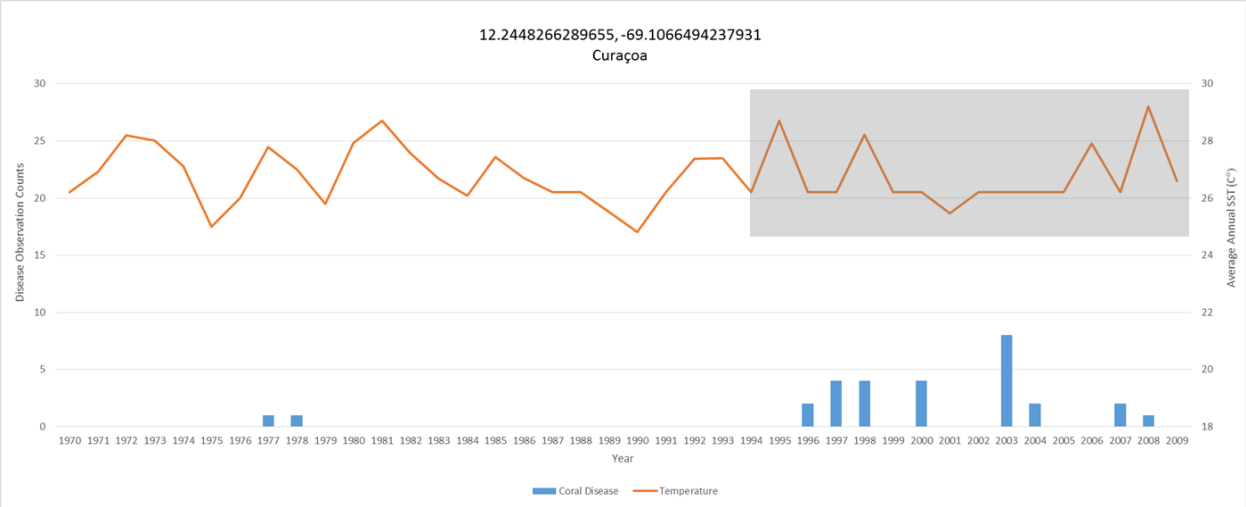

Fig. 4 Coral disease and temperature analysis performed at a geographic location near Curaçoa. The highlighted region here represents what may be an anti-pattern – a trend that is not consistent with any recognized hypotheses regading temperature and coral disease. More subtle factors may be involved here.

While the complete review of all 293 analysis artifacts is outside the scope of these proceedings, the cases examined above captures the core result of this research: the ability to make better use of the vast amounts of data available in driving hypothesis generation and testing. Introducing a data-driven approach allows us to locate trends in the data that are both consistent and inconsistent with current hypotheses which can then further substantiate accepted claims in driving policy and management decisions as well as facilitate retooling of hypotheses where the initial claims are brought into question by the data.

Discussion

Improvements in instrumentation during the 1950s provided accuracies sufficient to estimate large-scale density fields from sea surface to abyssal depths. As data accumulated, scientists recognized the need for an atlas of the ocean during the early 20th century. The advancement in salinity measurements, though, seemed to renew the call for a global integrated ocean atlas as was exemplified by collaborative oceanographic programs during the 1950s that resulted in several regional atlases for sectors of the Atlantic. Yet as of 1978, the catalogue of oceanographic atlases prepared by Stommel and Fieux contained no global ocean atlas (Levitus 1982).

Since then many information systems have come online, but have suffered (i.e., lacked longevity in the data for scientific inquiry) from the same problem encountered among all information systems: the misconception that uniting data sources and offering access to the resulting repository is the most important element of data management (De Geus, 1997). As proven countless times in industry, though, it is the overall management of the information ecology – the data governance, strategy for use, and presentation – that is most important in the success of an information system (Davenport, 1997).

Implicit in our information ecology framework are general solutions to several major detriments to data analysis in hypothesis testing. Such problems include: data presentation medium, changes and evolution of underlying data sets, handling of missing data, and overall data utility. Investing software resources at the data allows for resolution of these problems in a general and robust fashion, providing a solid foundation upon which algorithms that capture the elements of scientific inquiry can be quickly and easily implemented. The separation of concerns and subsequent generalization of data tasks in software then makes more resources available for addressing the looming questions facing coral reefs today.

We cannot find any previous system that provides analysis on this scale, viz., 2^{293} different combinations that are available for the scientist to examine within this limited set of constraints alone. Furthermore, the abstraction of scientific inquiry in software logic from the data ecology gives a robust platform to easily extend the algorithm to meet scientists' evolving requirements. As demonstrated in the visualization presented earlier, scientists can drill down to a level of granularity suited for the generation of individual hypothesis testing.

Continuing with this research, some key priorities for future work include the integration of additional data repositories as well as additional views of the WOD data available, introduction of purely statistical and computational search metrics for analysis of data trends, and expansion of our data-driven hypothesis generation approach to scale with additional available biotic and abiotic factors.

Acknowledgements

We would like to thank the entire ICRS community for the opportunity to present our research, NOAA for providing the data used in this project, and Indiana University for the continual support.

References

Boyer TP, Antonov JI, Baranova OK, Garcia HE, Johnson DR, Mishonov AV, O'Brien TD, Seidov D, Smolyar I, Zweng MM (2013) World ocean database. NOAA Atlas NESDIS 72: 209

Davenport TH, Prusak L (1997) Information Ecology: Mastering the Information and Knowledge Environment. Oxford University Press: 28-45

De Geus AP (2002) The Living Company: Habits for Survival in a Turbulent Business Environment. Harvard Business School Press: 1-12

GCDD (2016) Global Coral Disease Database. UNEP World Conservation Monitoring Centre. http://gcdd.tinypla.net/diseases

Jones RJ, Bowyer J, Hoegh-Guldberg O, Blackall LL (2004) Dynamics of a temperature-related coral disease outbreaks. Mar Ecol Prog Ser 281: 63-77

Levitus, S (1982) Climatological atlas of the world ocean. NOAA Prof. Pap. 13

Locarnini RA, et al (2013) World ocean atlas 2013, volume 1: Temperature. NOAA Atlas NESDIS 73: 40

McClanahan TR, Mebrahtu A, CA Muhando, J Maina, MS Mohammed (2007) Effects of climate and seawater temperature variation on coral bleaching and mortality. Ecol Mon 77(4): 503-525

Porter JW, Dustan P, Jaap WC, Patterson KL (2001) Patterns of spread of coral disease in the Florida keys. The Ecology and Etiology of Newly Emerging Marine Diseases: 1-24

WorldFish (2016) Coral Bleaching GIS Dataset. http://www.reefbase.org/gis_maps/datasets.aspx