

Inferring distributions of organisms from sparse occurrence data: KGSMapper

Daphne G. Fautin (fautin@ku.edu) and Robert W. Buddemeier
 Department of Ecology and Evolutionary Biology, and
 Natural History Museum, University of Kansas, Kansas, USA
 Kansas Geological Survey, Kansas, USA



The true distribution of most marine organisms is poorly known, but modern technology is making rapid inroads to improve knowledge in this critical realm. Relevant technologies include both direct and indirect methods of understanding marine organism distributions. We have developed

a tool to infer from existing distributional data where else individuals of particular species might occur.

The time-honoured method of determining what occurs where is by documenting occurrences in the field. Some modern technologies that allow such direct measurements without having to be in the field (including passive acoustics, image analysis systems, and telemetry) are not practical for the vast majority of marine species. This includes nearly all invertebrates, which make no noise, move little or not at all, are of little or no direct economic value, and may be very small; furthermore, individuals typically occur among, on, or in other organisms.

We and collaborators have developed the web resource "Biogeoinformatics of Hexacorals" (<http://www.kgs.ku.edu/Hexacorals>) that consists of biological and environmental data linked to each other and to a geographic information system. The biological component contains published occurrence records of sea anemones, corals, and their relatives. Directly accessible as "Hexacorallians of the World" (<http://geoportal.kgs.ku.edu/hexacorals/anemone2/index.cfm>), it is underlain by a relational database that includes bibliographic, taxonomic, and biogeographic data. The environmental data, drawn from public sources such as the World Ocean Atlas, currently include global coverage for 43 physical and chemical parameters gridded in register at a half-degree resolution.

Both the biological and environmental data are served to OBIS – the Ocean Biogeographic Information System – a growing inventory of occurrences of marine organisms described by Vanden Berghe (2007). OBIS passes the biological data on to GBIF – the Global Biodiversity Information Facility (<http://www.gbif.org>). These distributed systems rely on contributors making data publicly accessible, recognising that multiple sources of distributional data can provide a more accurate biogeographic picture than any one alone.

The biological and environmental data interact through KGSMapper, a tool that allows us to infer, based on known occurrences, where else individuals of a species might occur. It is one of the "software tools for data exploration and analysis" referred to by Vanden Berghe (2007). It resembles conceptually a growing number of such tools that rely on modern computing power applied to increasingly accessible, and increasingly rich,

Variable Name	Mean	Std. Dev.	One Std. Dev. Range	Two Std. Dev. Range	Entire Range	Use to Find Similar Areas	Use for upper link	Use for lower link
ARAGONITE SATURATION	1627.17	1018.15	1 to 2200.1	1 to 5218.43	1 to 7607	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ARAGONITE SATURATION	54.20	1203.47	1 to 1061.73	1 to 2164.2	1 to 5444.5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CHLOROPHYLL A CONCENTRATION	248.52	187.27	1 to 1001.58	1 to 1822.86	1 to 4766	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
TEMPERATURE	27.52	400.08	0 to 780.02	0 to 1554.71	0 to 3022.87	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 1. Display for some of the environmental parameters used by KGSMapper and boxes for selecting which are to be used in generating the model output.

biodiversity data (see Elith *et al.*, 2006 for a review). Accessible through "Hexacorallians of the World," OBIS, and FishBase (<http://www.fishbase.org>), KGSMapper provides a workspace into which anyone can import organism occurrence data for analysis, alone or in combination with data from the database through which it is entered.

KGSMapper displays the value of all 43 parameters for each half-degree cell in which an organism has been recorded (more than 43 values are shown because some of the datasets are divided temporally; e.g. values for chlorophyll a concentration are provided on an annual basis and in three month periods). For an organism that occurs in more than one cell, KGSMapper calculates the mean and standard deviation for the values of all parameters in all cells in which an occurrence is recorded. These data give an idea of some of the physico-chemical attributes of the environment suitable for the target organism. KGSMapper then searches for, and displays, all cells in the world having values matching those for places where members of the target species are known to occur. These are places where the habitat is most likely to be suitable for the species, whether or not it has been recorded from there.

Inferring undocumented occurrences by using attributes of the habitat where individuals are known to occur is common to all such biodiversity modelling tools. Aside from having been designed to operate in the marine environment (where some of the others perform poorly), KGSMapper is distinctive in input, processing, and output.

INPUT: The user controls which of the 43 parameters will be used to search for other cells that have similar values (Fig. 1). This allows the user to apply expert knowledge about which parameters may control a species' distribution (e.g. aragonite saturation) and which are likely to be irrelevant to it (e.g. oxygen saturation). Trying combinations of parameters allows exploration of the factors important to a species.

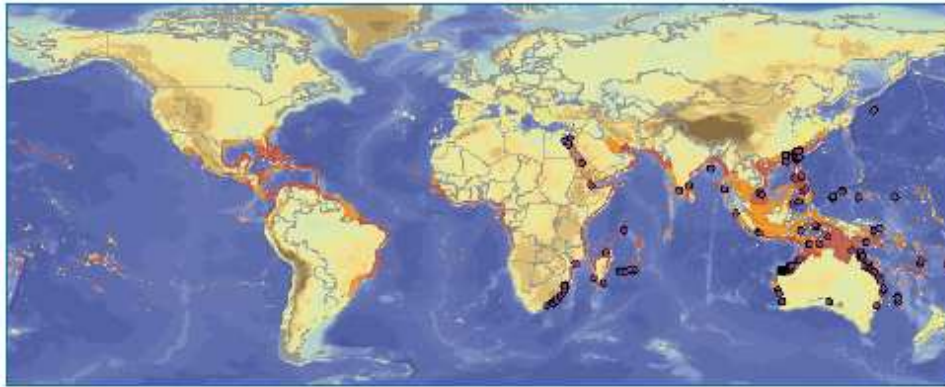


Figure 2. World map showing known occurrences of *Pterois volitans* (purple-ringed dots) and inferred suitable habitat (red and orange areas).

PROCESSING: The basis of the match is simple statistics. For each parameter, the algorithm calculates the mean and standard deviation of all cells containing at least one occurrence of the target species. It then finds all other cells with values within the 1-SD range for cells containing occurrences, within the 2-SD range, and in the entire range for all the parameters selected.

OUTPUT: All cells in which all parameters selected to generate the map are within the 1-SD range of the value of those parameters of cells in which the species is known to occur are displayed in red. Cells in which parameters outside of 1-SD are all within the 2-SD range are displayed in orange. Cells in which parameters are within the full range, but at least one is beyond the 2-SD range, are displayed in ochre.

The colour-coded output map serves as a rough measure of probability. It also serves as a means of data-cleaning. We have found that values beyond two standard deviations are often the result of a misplaced point or an anomalous habitat (for example, an intertidal organism might occur on a small island surrounded by deep water, so the average depth of the half-degree cell is not representative of the habitat of the target organism). KGSMapper can remove the values beyond 1 or 2 SD; we have found the red areas of a map very likely to provide a suitable habitat. Guinotte *et al.* (2006) described KGSMapper in detail, and assessed some facets of its accuracy and precision.

Expert knowledge is essential. Figure 2 is a world map showing the distribution, according to FishBase, of *Pterois volitans*, the Red Lionfish, in purple-ringed dots. Areas suitable for the occurrence of this fish – using the parameters of annual mean surface temperature, minimum monthly surface temperature, and minimum depth (constrained to depths no greater than 200 m; see Figure 1) are coloured red and orange (the ochre is faint and there is very little of it). Clearly the fish is endemic to the Indo-West Pacific. But inferred suitable habitats also exist in the eastern Pacific and Atlantic. Within the native range of the fish, the map can be used to infer where the fish might

occur other than where it has been reported (as in planning a fieldtrip, or assessing areas for conservation value). Outside that area, the map can be used to identify areas vulnerable to invasion. Most of the western Atlantic coast from southern Brazil to the central US appears to be suitable or very suitable for *P. volitans* with respect to the parameters used to generate the map. And, indeed, recently-established breeding populations of *P. volitans* have been reported from the central eastern United States (e.g. Semmens *et al.*, 2004).

Currently, KGSMapper is linked to environmental data for the surface (e.g. SST and chlorophyll *a* concentration) and the bottom (e.g. depth and bottom oxygen saturation). In theory, it could be used for the pelagic realm. Likewise, the grid size could be reduced to increase precision. We have demonstrated that KGSMapper is scale-independent by using it with finely-gridded data for the Hawaiian Islands (<http://hercules.kgs.ku.edu/hexacora/hawaii/biodata>). Extending KGSMapper to the pelagic realm and a finer grid would require significantly more environmental data and commensurate slowing of the process. But in future, with faster computers, such tools are certain to be developed.

References

Elith J., C.H. Graham, R.P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huetmann, J.R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J.M. Overton, A.T. Peterson, S.J. Phillips, K.S. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberón, S. Williams, M.S. Wisz and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29(2): 129-151.

Guinotte J.M., J.D. Bartley, A. Iqbal, D.G. Fautin and R.W. Buddemeier. 2006. Modeling habitat distribution from organism occurrences and environmental data: a case study using anemonefishes and their sea anemone hosts. *Marine Ecology Progress Series* 316: 269-283 (open access <http://www.int-res.com/abstracts/meps/v316>).

Semmens B.X., E.R. Buhle, A.K. Salomon and C.V. Pattengill-Semmens. 2004. A hot-spot of non-native marine fishes: evidence for the aquarium trade as an invasion pathway. *Marine Ecology Progress Series* 266: 239-244.

Vanden Bergh E. 2007. The Ocean Biogeographic Information System (OBIS). *GLOBEC International Newsletter* 13(2): p.83.