Automated Search for the Quantitative Laws Affecting CO_2 Fugacity in Sea Water

Kasun Wickramaratna, Miroslav Kubat
Dep't of Electrical and Computer Eng.
University of Miami
{k.wickramaratna@umiami.edu,mkubat@miami.edu}

Peter Minnett
Rosenstiel School of Marine and Atmospheric Science
University of Miami
{pminnett@rsmas.miami.edu}

Abstract

To describe and explain the world, scientists search for equations that quantify relations among relevant variables. We wanted to assist these efforts by a computer program inspired by genetic programing. Using real-world data, the program discovered equations that outperformed, in terms of accuracy, published expressions. Moreover, the research indicated that somewhat different formulas may be needed in different geographical regions. The latter observation appears to be due to "hidden" parameters that were not included in the available data.

Introduction

Our work was motivated by an important climatological problem: the question how CO_2-fugacity in the ocean depends on other variables, especially those measured from satellites. Earlier work relied on regression which produces standardized "black-box" formulas that cannot be easily interpreted. Moreover, we wish to see whether the prediction accuracy of the regression-based expressions can be improved. One limitation in existing algorithms is indicated by an observation made by (Olsen, Triñanes, & Wannikhof 2004) who noticed that the accuracy of linear regression improved when the set of independent variables was increased by geographic coordinates. In their formula, f_{CO_2} is CO_2 fugacity, T is sea surface temperature, l_t is latitude, and l_g is longitude:

$$f_{CO_2} = 10.18T + 0.53l_t + 0.292l_g + 52.2 \qquad (1)$$

Here, geographic coordinates seem to serve as proxies for other parameters that are either unknown or are only reflected by variables that are difficult to measure. Fortunately, we can expect the values of these unknown variables to be fairly stable over large geographic areas.

Relevant Work in Machine Learning

The problem of induction of equations that bind two or more variables was addressed by machine learning in the 1980s. The first major work introduced the system BACON (Langley 1981) that relied on two interleaved AI-searches, one in the space of equations, the other in the space of measured data. Among other things, the program re-discovered the ideal gas law. Some important results of further attempts along these lines were reported by (Zytkow 1987).

The main difficulty of the quest for numeric laws turned out to be the intractable size of the space of equations that may account for existing data (Falkenheimer & Michalski 1990). To overcome this problem, (Zembowitz & Zytkow 1992) preferred to focus on bivariate equations and let the user predefine preferred operators and functions; (Schaffer 1993) reduced the search space by the assumption of the existence of a small set of functional templates, and (Todorowski & Dzeroski 1997) suggested that the user might specify the search space by context-free grammars.

GA-Based Search for Formulas

Table 1 gives the pseudocode of the baseline version of our formula-seeking program. Here, each specimen represents one formula expressed as a tree-like data structure. The fitness function is defined as the formula's mean error on the training data. Initial population is a set of randomly generated formulas involving a predefined set of variables and operators. To these formulas, we "seeded" Equation 1 and a few formulas that we believed may steer the search in the right direction. Recombination was implemented as the exchange of random subtrees between a pair of formulas. Mutation affected variables, arithmetic operators, and coefficients. After the application of recombination and mutation, our program mixed the parents and children, arranged them by their error on training data, and then retained n best individuals.

To find N "regions of relative stability" (and the formulas corresponding to each of them), we modified the baseline algorithm as follows. Each specimen consists of N chromosomes (defining N different formulas) plus an $(N+1)$th chromosome that defines the boundaries of the given regions. To start with, we arranged the training examples by their latitude, and then divided them into N equally-sized regions. Then, we introduced an additional mutation operator to be applied only to the $(N+1)$th chromosome: for instance, if the second boundary is chosen, then a randomly generated integer, -9, changed the chromosome $[2,000; 4,000; 6,000; 8,000]$ to $[2,000; 3,991; 6,000; 8,000]$.

Table 1: Baseline version of the formula-searching system.

Until a termination criterion has been satisfied:

1. Randomly selected mating partners exchange genetic information by the recombination operator.
2. 20% randomly selected individuals are subjected to mutation of variables; 20% to mutation of operators; and all individuals to mutation of coefficients.
3. The fitness of each individual (children as well as parents) is obtained as the formula's error on training data. The top 50% are retained and the remaining formulas are discarded.

Experimental Results and Observations

We used data from `Explorer of the Seas` that transits the Caribbean once a week, taking measurements every 2 to 5 minutes (Williams, Prager, & Wilson 2002). Our program ran for 30 generations with the population of 500 specimens and $n = 5$ regions. From the initial equally-sized regions, the system developed the regions shown in Figure 1.

Figure 1: Regions identified by the genetic search.

Of the 500 different formulas found for each region, we removed those that contained latitude and longitude and those that were too complicated. Many formulas needed post-processing to remove suspicious terms (e.g., $\log(\log(\sin t))$), others could be improved by rules such as $\log(t + t) = \log(2t)$ or $\log(\exp(t)) = t$. To give the reader an idea of typical results, Table 2 compares, for each region, the error of a discovered formula with the one recommended by (Olsen, Triñanes, & Wannikhof 2004) (see Equation 1). p is surface pressure, and s is water salinity.

Conclusion

We have reported our experiments with a system capable of suggesting laws that may underly CO_2-fugacity. The endeavor was successful in the sense that the discovered formulas outperform equations put forward by our predecessors who used linear regression. Our most important finding is that the accuracy of fugacity predictions clearly improves if the system divides the data into geographic regions, and induces a different formula separately for each of them. We explain this observation by the existence of "hidden parameters" that vary in time and space, and, individually or in combination, exert different influence in different areas. The fact that our system seems to be able to discover the regions of their relative stability is encouraging.

Table 2: Selected equations for the individual regions

	Equation	Error
region 1	Olsen et al.	5.2 ± 4.8
	$13.36 * T$	4.0 ± 3.9
region 2	Olsen et al.	3.8 ± 3.1
	$56.0 + 11.1 * T + \log(T + p)$	3.7 ± 3.4
region 3	Olsen et al.	8.2 ± 6.2
	$390.0 + 0.33 * T^2 - 9.9 * T$	5.3 ± 4.1
region 4	Olsen et al.	9.0 ± 8.0
	$395.9 + 0.33 * T^2 - 9.9 * T$	5.5 ± 6.0
region 5	Olsen et al.	11.8 ± 7.4
	$\sqrt{(182.27 + s - T)} * T$	5.5 ± 5.1

From the perspective of machine-learning research, we would like to remind the reader of the circumstance that virtually all previous systems capable of numeric-law discovery only managed to re-discover previously known laws. Ours may be the first project that has contributed to a pressing environmental problem using such tools.

Acknowledgment

The research was supported by the NSF grant IIS-0513702.

References

Falkenheimer, B., and Michalski, R. 1990. Integrating quantitative and qualitative discovery in the ABACUS system. In Kodratoff, Y., and Michalski, R., eds., *Machine Learning:An AI Approach*. Morgan Kaufmann.

Langley, P. 1981. Data-driven discovery of physical laws. *Cognitive Science* 5:31–54.

Olsen, A.; Triñanes, J. A.; and Wannikhof, T. 2004. Air-sea flux of co_2 in the caribbean sea, estimated using in situ and remote sensing data. *Remote Sensing of Environment* 89:309–325.

Schaffer, C. 1993. Bivariate scientific function finding in a sampled , real-data testbed. *Machine Learning* 12:167–183.

Todorowski, L., and Dzeroski, S. 1997. Declarative bias in equation discovery. In *Proceedings of the 14th International Conference on Machine Learning*, 376–384. San Mateo, CA: Morgan Kaufmann.

Williams, E.; Prager, E.; and Wilson, D. 2002. Research combines with public outreach on a cruise ship. *EOC* 83(1-2):590–596.

Zembowitz, R., and Zytkow, J. 1992. Discovery of equations: Experimental evaluation of convergence. In *Proceedings of the 10th National Conference on Artificial Intelligence*, 101–117. San Mateo, CA: Morgan Kaufmann.

Zytkow, J. 1987. Combining many searches in the FAHRENHEIT discovery system. In *Proceedings of the 4th International Workshop on Machine Learning*, 281–287. Los Altos, CA: Morgan Kaufmann.