Mining for ecological thresholds and associations in cytometric data: a coastal management perspective

G. C. Pereira, A. R. Figueiredo & N. F. F. Ebecken
Federal University of Rio de Janeiro, COPPE/UFRJ, Brazil

Abstract

Decision-making in coastal waters management is a complex and interdisciplinary task. Particularly, to find seasonal patterns and ecological thresholds, which are not always clear in tropical areas. Therefore, the ultimate in this activity is to gain knowledge about biogenic element, the biological response, and the selection of indicators which may reveal the trophic status of the system. Under this scenario, this paper applies Data Mining techniques as an alternative approach in order to access hidden patterns of in situ flow cytometry monitoring data. The case studied is the upwelling influenced bay at Cabo Frio Island (Rio de Janeiro-Brazil). A neural network uses phytoplankton and bacterial data of real time monitoring as input variables to forecast marine viruses temporal variability. We also demonstrate that it is possible to access patterns of planktonic community structure in different water masses within a set of association rules.

Keywords: knowledge discovery, data mining, pattern recognition, environmental monitoring, coastal management.

1 Introduction

Pollution of coastal environment is a serious environmental problem and affects both developed and developing countries. It comes from different sources [1] causing mainly eutrophication of the coastal waters [2]. In this way, the Brazilian coast that presents a large variety of marine ecosystems and habitats are subjected to discharge of contaminants via sewage, industrial effluents, dredged material, accidental chemical and oil drilling spills, urban and agricultural runoff and atmospheric deposition from land-based activities like worldwide [3]. The

doi:10.2495/DATA090091

continuous reception of pollutants in coastal waters results in adverse effects on different organisms from the molecular level to the community level thus reducing the biodiversity and productivity of marine ecosystem and depletes marine resources [4].

In Brazil, the standards of environmental quality are still based in concepts of maximum admissible concentration level of pollutant according to the National Environmental Council (CONAMA). However, many authors have used phytoplankton [5, 6] or bacteria [7, 8] as bioindicators. However, viruses have been considered ecologically important members of aquatic communities as they influence biogeochemical cycles, community composition and horizontal gene transfer [9, 10]. In this way the ecological communities, plays a complex net of trophic interactions comprising many types of organism and should be used as a baseline indicator of ecological status through its biological integrity. Consequently, the management of coastal zone received more attention and efforts of different modelling approaches for predicting indicators as descriptors of system behavior [11–15].

Once a virus cannot be detected in real time monitoring, the aim of this work is to apply artificial neural networks (ANN) for estimating marine viruses from bacterial and phytoplankton abundance and association rules algorithm in order to find patterns of plankton community structure *in situ* and *ex situ* flow cytometry monitoring data.

2 Material and methods

2.1 Studied areas

The Southwest Atlantic Ocean off Brazil is known by its oligotrophy due to the prevailing Brazil Current (BC) that runs southwards, carrying Tropical Water (TW) from the vicinity of the Equator. Moving in the bottom on the opposite direction, there is the cold South Atlantic Central Water (SACW or ACAS) mass. In Arraial do Cabo, Northeast of Rio de Janeiro state (23°S, 42°W), the action of E-NE winds results in a shunting of the nutrient-depleted surface TW of Brazil Current to offshore followed by the up-flow of the deeper and nutrient-rich SACW. The inverse pattern comes with the S-SW winds when cold fronts bring the oligotrophic TW back to the coast.

2.2 Sampling procedure

Seawater samples were collected at the surface with a Nansen bottle with a reverse thermometer outside. Salinity, were determined ashore as described in [16]. Simultaneously, in situ flow cytometry measures were done. Samples of 200 ml were immediately fixed with 4% of paraformaldehyde for further cytometric (~3h) analyzes. Water masses identification and clustering were made according to the interval of temperature and salinity data provided by the Oceanography Department of Instituto de Estudos do Mar Almirante Paulo

Moreira-IEAPM (data not shown). The time series used for model predictions starts in August of 2006 and finished in June of 2007.

3 Flow cytometry

Phytoplankton enumerations were performed *in situ* with the CytoBuoy bench top flow cytometer [17]. Parameters were collected on a log scale using the CytoSift software and analyzed in the CytoClus software, both provided by the manufacturer (Figure 1a).

The real time enumeration of heterotrophic bacteria were performed simultaneously to the phytoplankton but the discrimination criteria was to filter data that have cytometric signatures with appropriate size (bacterial size range), no fluorescence (heterotrophy indication) and high side scatter signal (great metabolic activity-alive). This strategy was validated through comparisons with stained samples in the same cytometer in the lab.

The virus enumeration were performed with a FACScan flow cytometer (Becton Dickson, San Jose, Calif.) in the lab. Yellow-green 0,92-µm beads (Fluoresbrite Microparticles, Polysciences) were added to all samples as an internal standard (Figure 1b). The samples were stained with SYBR-Green-1 at a final concentration of 0.5 X 10^{-4} of the commercial stock solution according to [18]. The parameters for bacteria and viral counts were collected on a log scale and analyzed in the CellQuest™ Pro software provided by the manufacturer.

4 Analytical procedures

The statistical analysis and the ANNs models developed were performed in the Statistica 7.0 package software. For knowledge representation all variables were discretized into three intervals (low, mean, and high) according interviews with IEAPM plankton experts who set the cut points of variables. To this matrix was applied a modified Apriori algorithm in order to mine associations rules (if-them type) in the CBA 1.0 software [19].

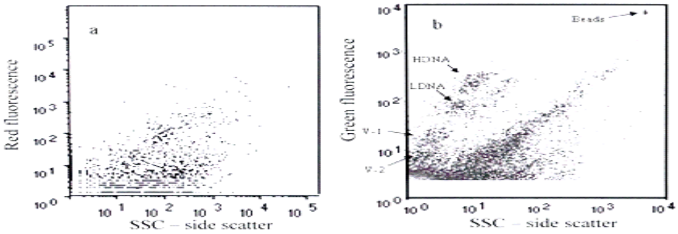

Figure 1: In a are present all phytoplankton distribution according their side scatter and chlorophyll fluorescence (red fluorescence) measured by the CytoBuoy cytometer. In b, one SYBR Green I stained sample demonstrating two groups of bacterial (LDNA, HDNA) and viruses (V-1, V-2) with different green fluorescence intensities respectively.

5 Results and discussion

Each dot depicted in Figure 1 is a suspended particle read by the two cytometers. In the case of Figure 1a, it presents the spread of real time data of phytoplankton cells acquired (radio transmitted) by the CytoBuoy flow cytometer since the red fluorescence signals are the results of chlorophyll-a response to laser excitation. The side scatter (SSC) is usually considered a measure of complexity. During the studied period, the total phytoplankton concentration varied from 8,66 x 10^2 cells/ml^{-1} in the summer to only 330 cells/ml^{-1} in the winter.

Figure 1b shows the picoplankton particle distribution. Two clusters of bacterial populations (LDNA and HDNA) with different green fluorescence intensities are easily noted indicating different nucleic acid content. The optical signatures of these groups were used for validation of real time data (in CytoBuoy acquisitions) and the total bacterioplankton varied from 1,24 x 10^6 cells/ml^{-1} in the summer to 1,51 x 10^3 cells/ml^{-1} in the winter.

Two viral populations are also noted: V-2 with lower fluorescence intensity is usually considered to be composed of bacteriophages [20] while V-1 is a diverse group that infects eukaryotes [21]. The total virioplankton community varied from 2,86 x 10^6 particles/ml^{-1} in summer to 6,21 x 10^5 particles/ml^{-1} in the same season. The correlations among biotic and abiotic variables are demonstrated in Table 1. The highest and most significant correlation is observed between heterotrophic bacteria (Bac Het) and the virus group V-2 (r^2 =0.97, n = 39, p< 0.05). Coincidently phytoplankton (Phyto) shows the same correlation to V-2. In the same way, V-1 presents a smaller but still high correlation to bacteria and phytoplankton (r^2 =0.91, n = 39, p< 0.05).

Table 1: Studied variables: V1 and V2 (viral sub-groups of low and high fluorescence), VT, total virus (sum of V1 and V2), Bac Het, heterotrophic bacteria, Fito, fitoplankton, Larvae, total of meroplankton larvae, Temp, temperature, Sal, salinity. Significance in bold numbers. Level of significance p < 0.05.

Variables	V1	V2	VT	Het Bac	Phyto	Larvae	Temp	Sal
V1	**1.00**	**0,88**	**0,64**	**0,91**	**0,91**	0,12	-0,13	0,03
V2		**1.00**	**0,72**	**0,97**	**0,97**	0,22	-0,14	0,13
VT			**1.00**	**0,76**	**0,76**	0,22	-0,09	0,26
Het Bac				**1.00**	**1.00**	0,09	-0,16	-0,08
Phyto					**1.00**	-0,29	-0,16	-0,08
Larvae						**1.00**	0,08	0,07
Temp							**1.00**	-0,15
Sal								**1.00**

The total virioplankton community (VT), the sum of V-1 and V-2, also shows a significant correlation to bacteria and microalgae communities (r^2 =0.76, n = 39, p< 0.05). Thus, our results suggest that both viral groups are equally active in bacteria and phytoplankton; however, to date we do not have any indication about phage infecting eukaryotic cells. Table 2 also shows that V-2 is highly correlated to V-1 (r^2 =0.88, n = 39, p< 0.05) and this is more correlated to VT than V-1 (r^2 =0.72, n = 39, p< 0.05).

The clustering procedure allowed us to verify the mean occurrence of the studied variables in each water mass. These data are shown in Table 2.

Table 2: Mean values of studied variables in different wares masses of Arraial do Cabo upwelling system.

WATER MASS		V-1/ml^{-1}	V-2/ml^{-1}	VT/ml^{-1}	Het Bac/ml^{-1}	Phyto/ml^{-1}	Bac/Phyto	VBR	Larvae	Temp	Sal
SACW	n=4	**1,43E+06**	**8,04E+03**	**2,23E+06**	8,06E+04	1,14E+02	7	**28**	0,15	18,00	35,55
COAST/TROP	n=22	8,19E+05	4,69E+03	1,29E+06	**2,63E+05**	**1,29E+02**	20	5	**0,30**	23,12	35,59
TROPICAL	n=12	9,56E+05	5,89E+03	1,55E+06	1,40E+05	6,41E+01	**21**	11	0,26	**22,43**	**36,06**

It is evident that V-1 is numerically dominant over V-2. The total virus community (VT) is more abundant than bacteria in at least one order of magnitude and bacteria are three times higher than phytoplankton. The highest mean value of virioplankton occurs in the SACW (n=4) while the highest mean values of bacteria, phytoplankton and meroplankton larvae occurs in the mixing of coastal and tropical waters (Cost/Trop, n=22). The cold water of SACW shows the smallest bacterial/phytoplankton ratio (Bac/Phyto) and the highest virus/bacterial ratio (VBR) indicating a great viral activity. This fact could explain why the more nutrient rich waters are not so productive.

From a management point of view, we used a supervised approach and genetic algorithm to evolve neural networks models to estimate the viral abundance (production controllers) from our field measurements of phytoplankton (primary producers) and heterotrophic bacteria (secondary production).

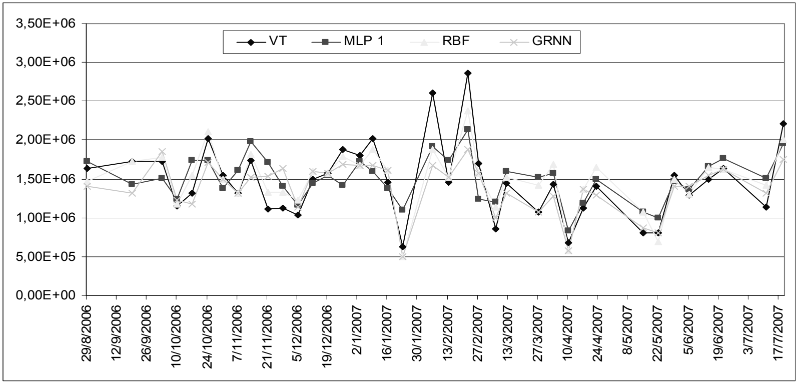

Figure 2: The three evolved neural network models behavior for total viruses (VT) prediction. MLP, multilayer perceptron; RBF, radial basis function and GRNN, general regression neural network.

Figure 2 demonstrates the behavior of three different type of ANN. Although the great variability of total virioplankton (VT = V-1 + V-2) shows two higher values in the middle of the studied period. These peaks are coincident with low temperatures due to upwelling events (not shown) in these date. In general all the three models are able to follow the behavior of this function however; the Multilayer Perceptron (MLP 1) neural network seems to be closer while Radial Basis Function (RBF) has a better estimate for higher values (peaks) and the General Regression neural network (GRNN) for smaller values.

Table 4 presents the RMSE error of training, test and validation data sets. It is clear that RBF is the more fitted model (R^2 = 0,97) to forecast viroplankton but these results should be viewed with some care due to RMSE penalize the error in the highest values.

Table 3: Neural network performance and error (RMSE) in training, test and validation data set.

Models	Train	Test	Validation	R^2
MLP	0,039232	0,029721	0,025814	0,84
RBF	0,005493	0,000427	0,000002	0,97
GRNN	0,005851	0,000923	0,000006	0,8

Although the huge natural variability of plankton we were able to mine and explicit some interesting association rules about the community structure. Examples are:

If TEMP low and SAL mean and PHYTO low them VIR high and BAC high [0.20%, 100%] (1)

If TEMP high and SAL mean and BAC mean them VIR mean [26.19%, 100%] (2)

If PHYTO high and BAC mean and TEMP mean them VIR mean [0.19%, 97%] (3)

If BAC high and VIR mean and SAL mean them PHYTO low [2.40%, 99%] (4)

If TEMP mean and BAC mean and PHYTO low them SAL high [7.35%, 99%] (5)

This set of rules reveals how community is structured and related to some environmental variables. It can be used as an initialization tool for visual inspections of food web interactions under abiotic conditions. The percentages between clasps mean firstly the support value which is the occurrence of the rule along the data set and secondly, its confidence level. The rule 1 is a clear example of SACW due to low temperature (TEMP) and salinity (SAL) is mean. Under this condition this rule depict also high virus (VIR) and low phytoplankton (PHYTO) values what was demonstrate in Table 2. The most common situation is present in rule 2 when Temperature increase to high and salinity decreased to mean. This is the case of the mixing of coastal and tropical water masses. In these case both bacteria and viruses shows mean value. We cannot forget although the interval mean, viruses are always the most abundant entity in the ecosystem. A special event of primary production is demonstrated in rule 3, in this case phytoplankton is high and bacteria are mean. It is example of mixing of SACW and the tropical water of Brazil Current. The rule 4 can be interpreted by another example of coastal and tropical water mixing, the difference of rule 2 are high bacteria values. Finally, rule 5 shows the oligotrophic condition of tropical water with low primary production of phytoplankton and mean values of bacteria occurring in the highest values of salinity and mean temperature.

5.1 Ecological interpretations

In the Arraial do Cabo upwelling system the emergence of SACW can find two different water mass, the Coastal or the Tropical. We hypothesize that upwelling events at least in narrow waters the up flow can bring mineral nutrient, organic matter, bacteria and viruses from the interface of sediment. These entrances of energy in water column take both bacteria and phytoplankton to growth but as they increase in abundance they also increase the susceptibility to virus infection. In this way, virus can control all the productivity in surface water but we speculate that some physical and biological factor can work at this moment. It is known that sum light can inactivate viruses and also as the water temperature increases and host availability decreases both factor synergistically working would induce virus populations to change from lytic to lysogenic life cycle. Thus, allowing the coexistence of all populations.

References

[1] Wu, J., Wang, J. 2008. Impacts of Pollution from Different Sources on Ecological Quality of a Multiple-use Coast. Water Air Soil Pollution 193:25–35.

[2] Eloffson, K., Folmer, H., Gren, I.M. 2003. Management of eutrophicated coastal ecosystems: a synopsis of the literature with emphasis on theory and methodology. Ecological Economics 47(1): 1-11.

[3] Pereira, G.C., Coutinho, R., Ebecken, N.F.F. 2008a. Data Mining for environmental Analysis and Diagnostic: a case study of Upwelling Ecosystem of Arraial do Cabo. Brazilian Journal of Oceanography 56(1): 1-12.

[4] Matthiessen, P., & Law, R. J. 2002. Contaminants and their effects on estuarine and coastal organisms in the United Kingdom in the late twentieth century. Environmental Pollution, 120(3), 739–757.

[5] Marshall, H.G., Lacouture, R.V., Claire Buchanan, C.B., Johnson, J.M. 2006._Phytoplankton assemblages associated with water quality and salinity regions in Chesapeake Bay, USA. Estuarine, Coastal and Shelf Science 69, issue 1-2, 10-18.

[6] Roelfsema, C.M., Phinn, S.R., Dennison, W.C., Dekker, A.G., Brando, V.E. 2006. Monitoring toxic cyanobacteria *Lyngbya majuscula* (Gomont) in Moreton Bay, Australia by integrating satellite image data and field mapping. Harmful Algae 5(1): 45-56.

[7] Noble, R.T., Moore, D.F., Leecaster, M.K., McGee, C.D., Weisberg, S.B. 2003. Comparison of total coliform, fecal coliform, and enterococcus

bacterial indicator response for ocean recreational water quality testing. Water research 37, 1637-1643.

[8] Frischer, M.E., Danforth, J.M., Foy, T.F., Jurasker. 2005. Bioluminescent Bacteria as Indicators of Chemical Contamination of Coastal Waters. Journal of Environmental Quality 34, 1328–1336.

[9] Weinbauer, M.G., Rassoulzadegan, F. 2004a. Are viruses driving microbial diversification and diversity? Environmental Microbiology, 6, 1-11.

[10] Suttle, C.A. 2005. Viruses in the sea. Nature 437, 356-361.

[11] Giacomini, M., Bertone, S., Caneva Soumetz, F., Ruggiero, C. 2005. An Advanced Approach Based on Artificial Neural Networks to Identify Environmental Bacteria. International Journal of Computational Intelligence 1(2): 90-97.

[12] Winter, C., Smit, A., Szoeke-Dénes, T., Herndl, G.J., Weinbauer, M.G. 2005. Modelling viral impact on bacterioplankton in the North Sea using artificial neural networks. Environmental Microbiology 7(6): 881–893.

[13] Dowd, M. 2006. A sequential Monte Carlo approach for marine ecological prediction. Environmetrics 17(5): 435-455.

[14] Mistri, M., Munari, C., Marchini, A., 2008. The fuzzy index of ecosystem integrity (FINE): a new index of environmental integrity for transitional ecosystems. Hydrobiologia (2008) 611:81–90

[15] Pereira, G.C., Ebecken, N.F.F. 2008b. Knowledge discovering for coastal waters classification. Expert Systems with Applications. Doi:10.1016/j.eswa.2008.10.009

[16] SCOR1996. Protocols for the Joint Global Ocean Flux Study (JGOFS) core measurements. Bergen, Norway: Scientific Committee on Ocean Research, International Council of Scientific Unions 9,170 p.

[17] Dubelaar GBJ, Gerritzen PL., 2000. CytoBuoy: A step forward towards using flow cytometry in operational oceanography. Scientia Marina 64(2): 255-265.

[18] Brussaard, C.P.D. 2004. Optimization of Procedures for Counting Viruses by Flow Cytometry. Applied and Environmental Microbiology, 70(3), 1506-1513.

[19] Liu, B., Hsu, W., Ma, Y. "Integrating Classification and Association Rule Mining." *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation)*, New York, USA, 1998.

[20] Weinbauer, M.G. 2004b. Ecology of procaryotic viruses. FEMS Microbiol. Rev. 28, 127-181.

[21] Larsen A., T Castberg, R-A Sandaa, C Brussaard, J Egge, M Heldal, A Paulino, R Thyrhaug, E van Hannen and G Bratbak. (2001). Population dynamics and diversity of phytoplankton, bacteria and virus in a seawater enclosure. Mar. Ecol. Prog. Ser. 221:47-57.