POLISH MARITIME RESEARCH Special Issue 2017 S2 (94) 2017 Vol. 24; pp. 89-94
10.1515/pomr-2017-0069

VISUALIZATION INVESTIGATION ON THE MARINE DATA WITH MULTIVARIATE STATISTICAL ANALYSIS METHODS

Li Yajie, Ph.D.,
Lv Zhengdong, B.S.,
Wang Maonan, B.S.,
Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

Marine information is an important way for us to know and study more about the ocean. Marine data makes the basic of marine information. Because of the huge quantity and diversity of marine data, and at the same time marine data is polyatomic variable, we start with statistical analysis methods to search for the regularity of the marine data. On one hand, we get the aggregate variation functions of the marine data by factor analyzing in aspect of the spatiality. Then we visually describe the marine status of the studied sea area with pre variogram function and post variogram function. On the other hand, we used cluster analysis method to get the verifying rule in time and make visible graphs of the marine data. In this way, we can also supply with the suggestions in classifying the sea seawater quality. The data processing result shows that the suggested methods in this article are both operable and effective. At the same time some reasonable suggestions are given in the article.

Keywords: marine data; factor analysis; cluster analysis, discriminate analysis; visualization

INTRODUCTION

The sea is closely related to a country's environmental protection, the resources developing and its national safety as well. And marine information is one of the important methods to know and study the sea [1]. Marine data is the basic of marine information. We can give important proves for studies in marine environmental monitoring, marine resources detecting and marine disasters forecasting [2]. Human's vision sense did great contribution to scientific studies. The conception of visualization was started to be used in the 80s in the 20th century. The massive of data could be converted into graphs in a visualization process. The graphs inspire our image thinking abilities which greatly improves data processing. Human's main visualization skills contain scalar field visualization, vector field visualization [3] and feature-based visualization [4, 5]. There are many marine data visualization studies in the world. For example, United States Naval Research Laboratory (NRL) made an data visualization research on the mixing process of the ocean in 1989. In the European eScience plan sponsored by the British National Institute for Environmental eScience (NIEeS) they started the research plan GODIVA [6] (Grid for Ocean Diagnostics，Interactive Visualization and Analysis). In this plan, they used the interactive visualization technology and network technology to process marine data. Poseidon [7], the oceanography study project was led by MIT. They used distributed computing as their basic method to integrate the marine data and made the model as well as functional modules, visualization and parameter calculation, etc. Visualization of marine data makes the ever changing massive marine data effectively used which helped to process and explain the marine data with high efficiency.

MATERIAL AND METHODS

CHARACTERISTICS OF THE MARINE DATA

Containing massive amount of information is a characteristic of marine data. The main method to collect marine data is to use remote sensing satellites and buoys. The seasats can help us to collect marine data by using the seasat sensors. Till 2012, more than 40 seasats have been sent to the out space by America, Australia and some other costal countries in the world. There are sea-viewing satellites for observing the seawater colours like Terra, Aqua and AMSE, etc. And sea-viewing satellites for observing the sea terrains like Topex, Poseidon, Jason-1/2, and CESat, etc. And also satellites for observing ocean dynamic like ERS-1/2, Envisat, HY-2, etc. The amount of the original marine data collected by the seasats is enormous. About 8500 agro buoys have been set on the sea by over 30 costal countries like US and Australia [8] which made a big sea monitoring network in the last 20 years. These buoys can collect the temperature information of the upper layer of the sea as well as some other data. Out from 1042 observing platforms of the National Data Buoy Center (NDBC) of America, 758 of them can supply real-time data [9].

Diversification is another character of the marine data. There are some main types of marine data like marine data collected by remote sensing seasats, hydrologic data, meteorological data, chemical data and biological data. And elements and forms contained by each type of marine data vary one to antoher. Marine chemical data contains dissolved oxygen, PH value, total alkalinity, active phosphorus, active silicate, acid salt, nitrate, nitrite, sulfide, organic pollution, and heavy metal and nutrient elements in forms like excel, mdb, csv, xml, etc. Another example proving that marine data varies in forms and types is that hydrologic data shows water temperature, height of tide, time of tide, salinity, wave number, wave height, water depth and transparency of the water, etc [10]. And if the data is collected by different monitoring instruments the data forms is different from each other. For example, we can see there are three types of form the marine data collected by remote sensing seasats: NUMBER (8) (20170408), String (10) (2016/04/08), NUMBER (4) (2) (2). Marine data has varies characteristics and types in different forms. This characteristics and types of marine data and data forms rely on each other and influences one another [11].

METHODS

Spatial is another attribute of the marine data and the data from different space have connections to each other. Marine data also has time attribute and data collected at different time are multivariate variable[12]. We want to do the study on the visualization work of the marine data in both the special aspect and the time aspect. At first, we use the simulated marine data to explain our working method and then do actual data analysis with real data. Let us say that the data in this simulate marine data are collected form N stations and there are P attributes C_1,C_2,…,C_p, in data from each station described in the following chart(here N and P stands number). See chart Tab.1:

Tab. 1. Marine Data Simulation

Station	C_1	C_2	…	C_p
A_1	C_{11}	C_{12}	…	C_{1p}
A_2	C_{21}	C_{22}	…	C_{2p}
⋮	⋮	⋮	⋮	⋮
A_n	C_{n1}	C_{n2}	…	C_{np}

SPECIAL ATTRIBUTE OF MARINE DATA

In order to inspect and study the special attribute of the marine data, let us say that the simulated marine data in table 1 was collected at certain time. This time could be a certain year, a season, a month, one day or an accurate time. We can set the time according to our studying targets. In this article we assume that we collected marine data in a certain year from N inspection stations and the data has P types of characteristics (here N and P stands for numbers).

If the sea is calm, then the data collected from the close stations would be relatively stable and tend to be similar. Then what are close stations? There are many characteristics, how can we define stable and similar? First of all, we need to find the close stations. We start from a station marked as $A_{(1)}$. Then we go to the closest stations marked as $A_{(i)}$. $A_{(2)}$ is the closest station to $A_{(1)}$. We mark closest station from A_i to $G_{(1)}$ (A and G are station types and $G_{(1)}$ is formed by $A_{(1)}$ and $A_{(2)}$) as $A_{(3)}$. Then we get the close stations $A_{(1)}$,$A_{(2)}$,…, $A_{(n)}$ to $A_{(1)}$. The distance to $A_{(1)}$ increases from $A_{(1)}$ to $A_{(n)}$.

In the discussion above, we talked about the distance between stations and between stations groups of different types. We get samples x_i and x_j. We mark the distance between the two samples i and j as d_{ij}. We have some types we station distance like Euclidean distance:

$$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^2 \right)^{1/2} \tag{1}$$

There are eight types of separation distances we often use in our esearches. We use G_p and G_q to mark two distance types. And the number of samples in each types marked as n_p and n_q. D_{pq} describes the distance between type G_p and type G_q. If we apply the method to describe the average distances between different types. The average distance between both types is the distance square in between:

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d_{ij}^2 \tag{2}$$

To make it easier to operate, we normally choose Euclidean distance as the distance between two stations and use the group average methods to describe the separation distance.

Now let's come to the second question: how to define and measure stable? On the first step, since there are marine data with P attributes (here P stands for number) in each station, there are large amount of attribute values, we descend he dimension of the variables with factor analysis method and measure the data stability of each station with general factor F. The work procedure is listed below:

Step I: standardize the original variables. We usually need to standardize the dimension of each attribute to make the data comparable. We mark the standardized attribute as C_{ij}. Then we evaluate for C_{ij}'s matrix R and its characteristic root w_i. The biggest characteristic root is marked as w_1 and the characteristic value decries from w_1.

Step II: evaluate the original common factor F and the factor load matrix with principal component analysis method.

Step III: Apply the orthogonal rotation method with maximized deviations to rotate the factor. Step IV: In this step we choose and explain the factors. Step V: Use the regression method to evaluate the factor score. See chart Tab .2.

In the real life, if we get the result that if the accumulated variance contribution rate :

$$\frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{n} w_i} \geq 70\% \tag{3}$$

We can say that we have collected the good original data. The value of common factor F is the weighted value of $F_1,\ldots,F_k$ which means $F=w_1F_1+\ldots+w_kF_k$. Then the weighted factor score of factor F we got from station $A_{(i)}$ is $F_{(i)}$.

Tab. 2. Simulated Marine Data in a certain year

Station	C_1	C_2	…	C_p	F
$A_{(1)}$	C_{11}	C_{12}	…	C_{1p}	$F_{(1)}$
$A_{(2)}$	C_{21}	C_{22}	…	C_{2p}	$F_{(2)}$
⋮	⋮	⋮	⋮	⋮	⋮
$A_{(n)}$	C_{n1}	C_{n2}	…	C_{np}	$F_{(n)}$

Then we measure the stability with the differential function of spatial value F (marked as FF). The calculation formula is:

$$FF_i = F_{(i)} - F_{(i-1)}, i = 2,\ldots,n \tag{4}$$

If the sea is calm, then the data collected from the close stations would be relatively stable and tend to be similar. This also means there should be no big difference between the value of FF (FF_i) we got at the beginning and the value of FF (FF_{i+1}) we got later. To make it visible, we put FF_i and FF_{i+1} into a scatter diagram. The coordinate of the first spot in the diagram is (FF_i ,FF_{i+1}). This idea was inspired by the scatter diagram electrocardiogram RR-Lorenz [13]. When the sample contains large amount of data, the scatter diagram shows like a baseball bat when the sea is calm. In this scatter diagram the line in center of the scatterplots has its gradient close to 1. We can also apply the regression analysis method to get the equation of the line for prediction purpose.

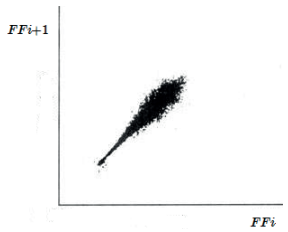

Fig. 1. FF-FF Scatter Diagram

TIME ATTRIBUTE OF MARINE DATA

We take the marine data collected from a same station for example. Let us use $A_{(1)}$ for the station. There are P attributes $C_1,C_2,\ldots,C_p$ in station $A_{(1)}$. In the data we use here we add the time that the data was collected as a part of the data. The collecting time could be a certain year, a month, a day or an accurate time. For example we can use days in our study. For example 1/1/2017 stands for mid day time of Jan. 1st 2017 as showed in Tab.3.

We can observe if there are clustering phenomena in the data we used in the study to the characteristics of time. We start with hierarchical clustering method to determine how many types we need. Say if we need k types of data, we need to apply K-means clustering method to do fast cluster analyzing to the data. We can classify G different types data into good, medium and low level according to the sea water quality standards and the central value classification we got from the K- means value clustering analyzing method .

Tab. 3. Simulate Marine Data from a same Station

Collecting Time	C_1	C_2	…	C_p	G
1/1/2010	C_{11}	C_{12}	…	C_{1p}	G_1
1/2/2010	C_{21}	C_{22}	…	C_{2p}	G_1
⋮	⋮	⋮	⋮	⋮	⋮
4/7/2017	C_{n1}	C_{n2}	…	C_{np}	G_2

Then we can make several types of statistical graphics like the Chernoff faces graph C_{ij} and the broken line graph of clustering value G, etc. of the collected data.

EMPIRICAL ANALYSIS

We choose some buoy data we quoted from the American National Data Buoy Centre (NDBC) website as our sample (see Fig. 2). Since the two types of buoys showed in triangle shape and diamond shape in the website stand for different types of data, we choose the data presented by the diamond shaped buoy in our article. We chose three groups of data

collected with buoys separately from the top, middle and the bottom. When the buoy is selected, it turns red. We can get the data we need when we input the information about time and data type by selecting in the web. Then we can get implications and scopes or each attribute by clicking *submit*.

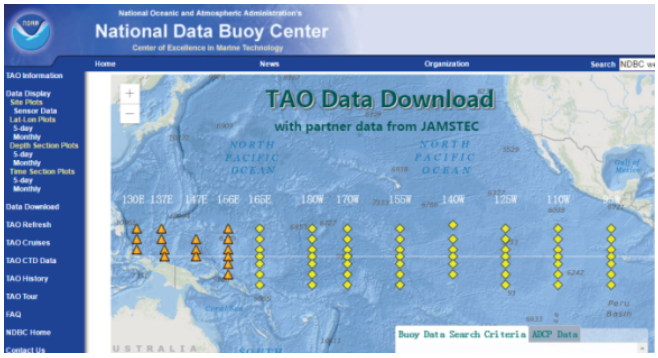

Fig. 2. Marine data form a sea area
Data source: http://tao.ndbc.noaa.gov/tao/data_download/search_map.shtml

When we choose 24 buoys as the sample data resources (considered location factor we chose T8N165E,T0N165E, T8S 165E,T8N180W,T0N180W,T8S180W,T8N170W,T0N170W,T8 S170W,T8N155W,T0N155W,T8S155W,T9N140W ,T0N140W ,T5S140W,T8N125W,T0N125W,T8S125W,T8N110W,T2S110 W,T8S110W,T8N95W,T0N95W,T8S95W).

We noticed that there are two buoys with big difference in attribute from the others so we deleted the related data and the sampled buoys reduced to 22(T0N140W,T0N170W). These buoys are marked as $A_{(1)}$,$A_{(2)}$…, $A_{(22)}$.

The observing time is from 1/1/2012 to 4/7/2017. We got the data showing 9 attributes (Surface – Met, Wind, Relative Humidity, Air temperature, Sea surface Temp, Subsurface temp, Sea water pressure, Salinity, Density). And in this data pack there are at most 46 sub-attributes. The data form got from the website is ASCII (American Standard Code for Information Interchange). In order to do the study in this article, we converted the data form into CSV (compatible of Excel).

Marine data has the characteristic of huge amount and multiformity, it also varies in time and space. Each buoy (station) has the information of time, space, hydrological attribute elements and the monitored marine data. We processed the missing information (with near interpolation method and mean value interpolation method) and sorted the data characteristics collected from different stations. As a result, 23 sub-attributes including time, AIRT, SSD, UWND, VWND, WSPD, WDIR, AIRT, RH, PRES, PRES, SSS, SST and TEMP, etc are analyzed in the process. The total value quantity of this data was over 30 thousand. A part of the values are shown in the chart below:

Tab. 4. Chart for the processed data of marine data from a sea area

fb	YYYYMMDD	HHMMSS	AIRT (hei	SSD (dept	UWND (hei	VWND (hei	WSPD (hei	WDIR (hei	AIRT (hei	RH (heig	PRES (dep	PRES (dep
T8N165E	20120102	120000	28.38	-99.9	-7.3	-2.4	7.7	252.2	28.38	82.5	300.624	503.046
T8N165E	20120103	120000	27.16	-99.9	-4.9	0.4	4.9	275.2	27.16	87.5	300.395	502.513
T8N165E	20120104	120000	28.11	-99.9	-5.4	-1.9	5.7	250.9	28.11	81.8	301.128	503.989
T8N165E	20120105	120000	28.11	-99.9	-8.1	-2.6	8.6	252.1	28.11	80.2	298.938	499.127
T8N165E	20120106	120000	27.81	-99.9	-7.2	-2.8	7.7	249	27.81	76	299.565	500.256
T8N165E	20120107	120000	27.66	-99.9	-7	-1.4	7.2	258.6	27.66	83.7	300.468	502.419
T8N165E	20120108	120000	26.94	-99.9	-7.9	-2.3	8.2	253.9	26.94	90.9	298.784	498.684
T8N165E	20120109	120000	26.87	-99.9	-7.6	-2.7	8.1	250.3	26.87	96.7	298.75	498.703
T8N165E	20120110	120000	27.97	-99.9	-8.4	-2.8	8.8	251.5	27.97	80.6	299.659	500.588
T8N165E	20120111	120000	27.88	-99.9	-7.7	-3.3	8.3	246.8	27.88	76.8	299.692	500.754
T8N165E	20120112	120000	27.17	-99.9	-7.4	-3.1	8.1	247.2	27.17	82.5	299.297	499.72
T8N165E	20120113	120000	27.87	-99.9	-7.7	-1.9	7.9	256	27.87	75.6	299.699	500.55
T8N165E	20120114	120000	27.59	-99.9	-5.2	-1.7	5.4	251.9	27.59	71	300.57	502.491
T8N165E	20120115	120000	27.45	-99.9	-4.3	-0.8	4.4	259.1	27.45	72.6	300.396	502.011
T8N165E	20120116	120000	27.48	-99.9	-5.3	-3.2	6.2	238.8	27.48	74.1	300.128	501.548
T8N165E	20120117	120000	27.52	-99.9	-4.4	-3.2	5.4	234	27.52	77.5	300.604	502.455
T8N165E	20120118	120000	27.61	-99.9	-4.8	-3.1	5.7	236.7	27.61	83.7	300.165	501.457
T8N165E	20120119	120000	26.53	-99.9	-5.5	-2.8	6.2	242.8	26.53	93.6	299.941	501.161
T8N165E	20120120	120000	27.76	-99.9	-6.6	-2.5	7	249.5	27.76	79.5	298.97	498.926
T8N165E	20120121	120000	27.95	-99.9	-7	0.3	7	272.7	27.95	82.1	298.802	498.704
T8N165E	20120122	120000	28.09	-99.9	-2.8	0.4	2.8	277.5	28.09	79.5	301.022	503.472
T8N165E	20120123	120000	27.89	-99.9	-5	-3.3	6	236.5	27.89	81.4	300.486	502.586
T8N165E	20120124	120000	28.06	-99.9	-7.3	-2.4	7.7	252.2	28.06	78.7	298.46	498.131
T8N165E	20120125	120000	28	-99.9	-8	-3.4	8.7	247	28	74.5	300.068	501.496

We can choose the data collected in a certain day (eg. 9 19th, 2016) to do the factor analysis. We can also choose the mean values of data collected in a certain month or a year to do the data analysis. According to the variance contribution rates of the factors, the first five common factors' variance contribution rates reach 90.3%.

Value of general factor F is the weighted value of F_1,…,F_k. We get the value of F (FF) in through the equation :

$$F = 9.889F_1 + ... + 1.037F_5 \quad \textbf{(5)}$$

Then we get the result in the chart below by calculating the value of F with formula FF.

Tab. 5. FF Values

fb	F	FF
T8N165E	5.41	
T0N165E	45.52	40.11
T8S165E	-0.73	-46.25
T8N180W	1.11	1.84
T0N180W	-0.67	-1.78
T8S180W	-0.73	-0.06
T8N170W	0.37	1.1
T8S170W	-1.18	-1.55
T8N155W	-0.61	0.57
T0N155W	-6.52	-5.91
T8S155W	-3.93	2.59
T9N140W	6.43	10.36
T5S140W	-3.83	-10.26
T8N125W	5.84	9.67
T0N125W	-8.57	-14.41
T8S125W	-5.39	3.18
T8N110W	-4.8	0.59
T2S110W	-9.78	-4.98
T8S110W	-6.66	3.12
T8N95W	-4.93	1.73
T0N95W	-6.37	-1.44

We made Fig. 4 by making the scatter graph of the value of FF （FF_i） we got at the beginning and the value of FF（FF_{i+1}） we got afterwards. The reason might be the small quantity of station data we analyzed or there were some other changes in the sea area that day. It needs more specific analysis and study by the local inspecting people to make it clearer.

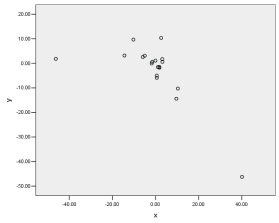

Fig.4. Scatter graph of the value of FF (FFi) got earlier and afterwards

We choose the data from a certain monitoring station (like T8N165E) to do cluster analysis according to the time line. The sampled time length is 1488 days from 1/1/2012 to 4/8/2017. After cluster analyzing, we found that the data should be sorted into three time periods. Then we applied k-mean value clustering analysis method and got the data process result of each station. According to the analyzing result, we made the time-cluster analyzing line chart (Fig. 5) below.

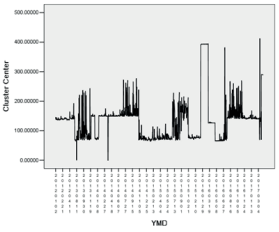

Fig.5. Time-cluster Analyzing line chart

We can see that the station has different conditions in different time periods. We need to pay more attention to the time spot appeared in the graph.

We can also sort the seawater quality into tree levels according to the seawater quality standard and the medium value of the cluster analyzing results thus to estimate the water quality in each time period. Level one suit for protecting the marine biological resources and the safety for human to use. Level two suit for bathing beaches and scenic spots. Lever three suit for developing industries, harbour water, and some other sea developing working area[14,15].

RESULTS

The visualization study of the marine data is of critical importance due to the huge quantity, it's varied attributes and multi-resolution of it[16]. Building the ocean model and the visualized model of marine data has become a very important research subject in studying the digital ocean.

DISCUSSION

We did some supplements to the original data when we realized there was so much missing information in the sampled data. This also reminds people to check if there are any malfunctions on the buoy, and if so, it calls for timely maintenance.

In this article, we did cluster analysis to the marine data due to its time attributes. We can do further analysis to the data with orderly cluster method to determine the rules since time has its continuity. After clustering, we can make prediction with discriminant analyzing method. Better effects could be achieved through combined effort of analyzing according to time and space factors which is in our future work. In this article we did the analysis separately.

ACKNOWLEDGEMENTS

This research is supported by National Natural Science Foundation of China under Grant No. 61375066.

BIBLIOGRAPHY

1. Smirnov G.V. ,Olenin A.L.: *Marine information systems and new measuring channels for hydrophysical parameters*, Oceanology, Vol.55, pp.291-295, 2015.

2. Andreas M. , Gunther R. : *Review of three-dimensional ecological modeling related to the North Sea shelf system. Part 1: models and their results*,Progress in Oceanography, Vol.57, no.2, pp. 175-217,2015.

3. Helman J, Hesselink L.: *Representation and Display of Vector Field Topology in Fluid Flow Data Sets*, IEEE Computer, Vol.22, no.8, pp. 27-36,1989.

4. Leeuw W. D., Liere R. V.: *Multi-level topology for flow visualization*, Computers & Graphics, Vol.24, no.3, pp.325-331,2000.

5. Reinders F., Post F.H., Spoelder H.J.W.: *Visualization of time-dependent data with feature tracking and event detection*,The Visual Computer, Vol.17, no.1, pp.55-71,2001.

6. Marsh J., Glencross M., Pettifer S., and Hubbold R.: *A network architecture supporting consistent rich behaviour in collaborative interactive applications*, IEEE Transactions on Visualization and Computer Graphics, Vol.12, no.3, pp.405-416,2006.

7. Evangelinos C., Lermusiaux P. F. J., Geiger S. K., et al.: *Web-enabled configuration and control of legacy codes An application to ocean modeling*, Ocean Modelling, Vol.13, no.3, pp.197-220, 2006.

8. Dai H.L., Mou N., Wang C.Y., et al. : *Development status and trend of ocean buoy in China*, Meteorological Hydrological and Marine Instruments (in Chinese), Vol.2, 2014.

9. Zhang Feng, Li Sihai, Shi Suixiang: *Research of Data Architecture in Digital Ocean*, MarJne Science Bulletin, Vol.12, pp.85-96,2010.

10. Levitus S., Antonov J. I., Boyer T.P., et al.: *World ocean heat content and thermosteric sea level change (0-2000m) 1955-2010*, Geophysical Research Letters, Vol.39, no.10, pp.L10603-L10607,2012.

11. Cummings J.: *Operational multivariate ocean data assimilation*, Quarterly Journal of the Royal Meteorological Society, Vol. 131, issue 613, pp.3583-3604, 2005.

12. Chau K., Muttil N.: Data mining and multivariate statistical analysis for ecological system in coastal waters. Journal of Hydroinformatics, Vol.9, no.4, pp.305-317,2007.

13. Chuang, S.S., Wu, K.T., Lin, C.Y. et al.: *Poincaré plot analysis of autocorrelation function of RR intervals in patients with acute myocardial infarction*, J Clin Monit Comput, Vol.28, pp.387-401,2014.

14. Hatzikos E., Hätönen J., Bassiliades N., Vlahavas, I., Fournou, E.: *Applying adaptive prediction to sea-water quality measurements*, Expert Systems with Applications, Vol.36, pp.6773-6779,2009.

15. Shrestha S., Kazama F.: *Assessment of surface water quality multivariate statistical techniques: a case study of the Fuji river basin*, Japan Environmental Modelling and Software, Vol.22, pp. 464-475,2007.

16. M. J. Martin, M. Balmaseda, L. Bertino, et al.: *Status and future of data assimilation in operational oceanography*, J. Oper. Oceanogr. , Vol.8, no.S1, pp.28-48,2015.

CONTACT WITH THE AUTHOR

Li Yajie

e-mail: lyj7712@163.com
tel.: 86-10-62282322
Beijing University of Posts and Telecommunications
Beijing, 100876
China