Estuarine, Coastal and Shelf Science 149 (2014) 255–263

Contents lists available at ScienceDirect

Estuarine, Coastal and Shelf Science

journal homepage: www.elsevier.com/locate/ecss

Prediction of near-bottom water salinity in the Baltic Sea using Ordinary Least Squares and Geographically Weighted Regression models

Katarzyna Łukawska-Matuszewska [a, *], Jacek Andrzej Urbański [b]

[a] Institute of Oceanography, University of Gdansk, Al. Pilsudskiego 46, 81-378 Gdynia, Poland
[b] GIS Centre, University of Gdansk, Bazynskiego Street 4, Gdansk 80-952, Poland

ARTICLE INFO

Article history:
Received 30 June 2014
Accepted 3 September 2014
Available online 16 September 2014

Keywords:
Baltic Sea
salinity
geostatistical modelling
Geographically Weighted Regression

ABSTRACT

A map of spatial salinity distribution in the bottom water layers of the Baltic Sea is presented in this paper. The map has been constructed based on the data obtained from the ICES Dataset on Ocean Hydrography. The typical salinity values and the depth of halocline location in the major basins of the Baltic Sea are also presented.

While the spatial salinity distribution is commonly derived by interpolation from the available data set, the linear regression model has been applied in this work. The analyzed data cover the period between 1913 and 2011, with a spatial resolution of ca. 10 km. In order to prepare the salinity map for the bottom water layers in the Baltic, the relationships between the salinity, depth and the distance from the Danish Straits have been determined by using Geographically Weighted Regression (GWR). Next, the salinity map was created by using the maps of regression coefficients, the digital elevation model (DEM) of the Baltic Sea, and the map of Euclidean distance from the Danish Straits. Subsequently the salinity values in the water layer above and below the halocline that are typical for the specific Baltic basins as well as the depth of location of the halocline were calculated based on the data extracted from the map by random point sampling.

The calculated salinity values for the upper layer were similar to the values reported in the current publications on the subject of the Baltic Sea. On the other hand, the obtained salinity values for the layer below the halocline were slightly lower than those found in the literature, which is attributable to different methodology used. The obtained results demonstrate that GWR is a reliable tool for estimating the natural variation of salinity in the Baltic Sea. At the same time, we conclude that the Ordinary Least Squares regression should not be used to analyze similar data.

1. Introduction

Baltic is a small (surface area of ca. $393{\cdot}10^3$ km^2), shallow (mean depth of 54 m) and semi-enclosed sea, which consists of a number of basins separated by the underwater sills (Leppäranta and Myrberg, 2009). The exchange of water masses between the North Sea and the Baltic, which occurs through the narrow and shallow straits (mainly Öresund and Great Belt), and the input of freshwater via riverine inflow and precipitation play the major role in shaping the salinity regime of the Baltic Sea (Döös et al., 2004; Reissmann et al., 2009). Deep water layers mainly originate from the so-called Major Baltic Inflows (MBI), which are strong enough to renew the bottom waters in the Baltic deeps (Döös et al., 2004; Reissmann et al., 2009). The frequency of MBI has decreased over the recent several years. At present, MBI occur irregularly and, on average, once in 10 years (Matthäus and Franck, 1992; Feistel et al., 2003; Matthäus, 2006; Meier et al., 2006). More saline waters flowing into the Baltic from the North Sea sink and fill the deep basins. These denser water masses are separated from the surface waters by a halocline, which occurs at the depth of 35–40 m and 70–90 m in the Arkona Basin and Gotland Basin, respectively (Reissmann et al., 2009). With increasing distance from the Danish Straits the oceanic water masses become more diluted. Moreover, the exchange of water among the specific Baltic basins is horizontally restricted due to the bottom topography (Döös et al., 2004). As a result, salinity in the bottom layers deceases with increasing distance from the straits (Bock, 1971).

* Corresponding author.
E-mail address: k.lukawska@ug.edu.pl (K. Łukawska-Matuszewska).

http://dx.doi.org/10.1016/j.ecss.2014.09.003
0272-7714/

The typical salinity values reported in the published literature for the different Baltic Sea basins (e.g. Leppäranta and Myrberg, 2009) date from the 1970s (Bock, 1971). The aim of this work was to create a map of spatial salinity distribution for the bottom layers of the Baltic Sea by using the data from the period 1913–2011, which had been obtained from the ICES Dataset on Ocean Hydrography (ICES, 2013). Besides the map, the updated salinity values in the layer above and below the halocline as well as the depth of halocline location in the main Baltic Sea basins are also presented. Moreover, different spatial regression techniques, i.e. Ordinary Least Squares Regression providing a global model and Geographically Weighted Regression providing a local model were compared with regard to their applicability for predicting salinity in the near-bottom water layer.

2. Materials and methods

2.1. Regression models

Statistical models used to make spatial predictions that are based on environmental factors can be classified into several groups (Breiman et al., 1984; Chambers and Hastie, 1992; McBratney et al., 2003; Bishop and Minasny, 2005; Hengl, 2009), as follows: (1) classification-based models (the outputs are class boundaries or classification rules), (2) tree-based models (the output is a decision tree), and (3) regression models (the output is a set of regression coefficients). The spatial linear relationships are examined with the use of regression models which allow not only to understand and explain, but also to predict the complex phenomena.

In this study the spatial relationships between the dependent and explanatory variables were modeled using Spatial Statistics Tools in ArcGis10. ArcGIS provides several tools for regression analysis, *inter alia*, Ordinary Least Squares (OLS) regression, which is the most known of all regression techniques. OLS provides a global model of the analyzed variables. The output of the analysis is a single regression equation describing the relationship between the dependent and explanatory variables across the whole study area:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \quad \text{where} \tag{1.1}$$

Y_i is a dependent variable; $X_{i1} - X_{i,p-1}$ are explanatory variables; $\beta_0 - \beta_{p-1}$ are the regression coefficients; and ε_i is the random error term (residuals).

With spatial data, the regression coefficients often do not remain fixed over space (Brunsdon et al., 1998; Fotheringham et al., 2002). The spatial regression technique suitable for exploring non-stationary relationships between the analyzed parameters is Geographically Weighted Regression (GWR). GWR produces a set of coefficients for each feature in the dataset instead of a single regression equation for the entire analyzed area. For every feature in the dataset, separate equations are constructed, which incorporate the dependent and explanatory variables of features falling within the bandwidth of each target feature (Cleveland and Devlin, 1988; Brunsdon et al., 1996, 1998). In the present study GWR was the second technique applied for predicting of spatial distribution of salinity.

2.2. The assessment of model performance

A model never fully imitates the reality. Based on the empirical data, it is only possible to closely approximate the reality. The overall model performance is measured with multiple R-Squared (R^2) and adjusted R-Squared (adjR^2) values, both ranging from 0 to 1. The adjR^2 reflects model complexity and is considered a more accurate measure of the model performance. The magnitude of residuals, i.e. the differences between the observed and predicted values of dependent variable, is another measure of a model fit; the smaller the residuals, the better fit of the model (Dodge, 2008; Fahrmeir et al., 2013; Kuhn and Johnson, 2013).

A widely used diagnostic tool for assessing the model performance is the Akaike's information criterion (*AIC*) (Akaike, 1973, 1983; Burnham and Anderson, 2004). *AIC* is used to compare the performance of models with different sets of independent variables or to compare the global (OLS) and local (GWR) models (Burnham and Anderson, 2002). The model with smaller *AIC* is considered to be more suitable.

A well-specified model includes all important explanatory variables to adequately represent the dependent variable (Burnham and Anderson, 2002). When a model is incomplete the residuals show significant spatial autocorrelation. On the other hand, the explanatory variables included in the model should not be redundant. Multicollinearity can, *inter alia*, increase the estimates of the parameter variance or generate models in which no variable is statistically significant even though R^2 is large (Belsley et al., 1980; Greene, 1993; O'Brien, 2007). One of the methods for determining the presence of multicollinearity is the Variance Inflation Factor (*VIF*). The *VIF* tells us how much the variance of a coefficient associated with the explanatory variable increases because of the linear dependence between the explanatory variables. The variables associated with the high *VIF* values are usually eliminated from the model.

The regression coefficients computed for each explanatory variable represent the strength and type (positive or negative) of the relationship between the explanatory and dependent variables. The statistical significance of coefficients associated with each independent variable is assessed by *t* test. The coefficients with small *p*-values are important for the model, while the associated variables are effective predictors (Dodge, 2008; Fahrmeir et al., 2013; Kuhn and Johnson, 2013).

The analysis of model residuals provide important information about the bias of the model. In a properly specified model, the over- and underpredictions are normally distributed. The normality, homoscedasticity and the absence of autocorrelation in the residuals obtained from the linear regression models can be tested with the Jarque-Bera statistic (*J-B*) (Jarque and Bera, 1980). The statistically significant *J-B* statistic means that model predictions are biased. It is a sign of model misspecification or nonlinear relationship between one or more explanatory and dependent variables.

The OLS model is a global model under the assumption that the explanatory variables used in a model have a consistent relationship to the dependent variable within the studied area (the processes are stationary). If so, the variation of the relationship between the predicted values and explanatory variables does not change with their magnitude (there is no heteroscedasticity). In a linear regression model, the Koenker's studentized Bruesch-Pagan statistic (*K*(*BP*)) is used to test for heteroscedasticity. It tests whether the estimated variance of regression residuals is dependent on the values of independent variables (Breusch and Pagan, 1979). Assuming a 95% confidence level, the *p*-value smaller than 0.05 indicates the non-consistence (either due to heteroskedasticity or non-stationarity) of modeled relationships. When the relationships modeled with OLS are non-stationary one can include an additional variable in the model that explains the regional variation or use GWR, which does account for the regional variation.