D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

INTEGRATED USE OF SPATIAL DATA AND LEARNING ALGORITHMS TO DETECT WATER QUALITY TRENDS

Regina T. Kishi
Stephan Fuchs
Hermann H. Hahn
Institut für Siedlungswasserwirtschaft - ISWW
Universität Karlsruhe
Am Fasanengarten
76128 Karlsruhe – Germany
E-mail: kishi@iswws1.bau-verm.uni-karlsruhe.de

ABSTRACT

This paper presents the first results of a study which investigates river basin water quality management based on the ability of GIS to integrate different data sources into a common geographical database. The overall objective of this research is to develop a scheme for water management and improve the likelihood of detecting water quality trends and to determine if variability in water quality parameters among a number of catchments can be explained by their differences in topography, land use, soils and demographic data. The expected benefits of the research are related to the acquisition of a description of the water quality in basins using easily available data, as satellite imagery or maps, that can be used in regions where adequate observations are not available.

1 INTRODUCTION

One of the main tasks of water pollution control is to assess temporal and spatial variations of water quality and the factors and processes which influence it. Data and information about the basin, current land-uses, potential pollution sources, as well as its localization are essential for the management.

Several processes involved in pollution transport are complex and depend on a wide range of combined phenomena. The combination of different factors increase the potential of a pollution source. For example, erosion rates depend not only on the soil type, but on the combined effect of soil, slope, soil coverage and rainfall intensity.

The understanding of environmental processes demands appropriated information and data, but its availability is still one of the main problems to be solved. Most of the hydrologic variables vary along the basin and their analysis demands the processing of great amount of spatial data.

Current developments in the fields of digital information processing, automation of map recognition and spatial data capture make the use of more detailed data easier and more attractive, introducing changes in the development of water quality studies. The ability of GIS to integrate different spatial data layers in a common geographic database is a powerful tool in this sense.

In the present paper is described an approach for regional water quality assessment that compensates the lack of field observations using information derived from current available sources, as satellite imagery or digital terrain models. Existing data are processed within a GIS environment for the computation of parameters that describe the spatial variation of significant variables within the basin. These parameters are later used to feed a water quality model.

2 METHODS

2.1 Overview of the method

Pollution loads (dissolved or suspended) carried into streams depend on factors like land cover, soil use and meteorological, geological and hydrological characteristics of the basin. Therefore, an adequate description of their spatial variation provides clues about the water quality within the basin. Many studies report on the relationship between the chemical composition of streams and catchment characteristics (Jordan et al., 1997; Dillon & Molot, 1997; Wolock et al., 1990). Part of the necessary information does not experience great temporal variations and can be obtained from available maps, others, like land-cover, can be estimated from remote sensing imagery. There is also the possibility to integrate other information sources like demographic and economical data, available in form of official statistical reports.

In this approach, contaminant transport is described as a function of spatially referenced land-surface and stream-channel characteristics. For this purpose we analysed basin attributes like land-use or topography to estimate trends in water quality.

The Group Method of Data Handling (GMDH) algorithm was used to perform the information fusion and study its correlation with water quality observations. This kind of model is characterised by its ability to select the more suitable combination within the input variables set, using an algorithm that resembles the natural selection principle. In each iteration, the weakest variables are discarded and substitutes are generated as combination of the strong ones, in form of polynomials of second order.

Nitrate was chosen to perform the first set of tests. Annually averages of nitrate concentrations in rivers were considered. Nitrate is highly soluble and easily leached from soil, because of this the transport of nitrate takes place principally by sub-surface and groundwater flow. Earlier works showed that the principal sources of nitrate in surface waters are agricultural drain water, raw and treated sewage discharges, industrial discharges, pastures, livestock feed lots and landscaped urban areas (Beudert, 1997). Currently, the possible harmful effects of livestock on the surface waters and groundwater is being studied in the USA (Cressie, 1997). The same author reports that the EPA (U.S. Environmental Protection Agency) attempts to control pollution from feedlots with concentrated animal since 1972.

2.2 Study area

The study area is the Neckar river basin in Baden-Württemberg – Germany. Its total drainage area is approximately 14000 km². The basin of Neckar river covers a significant part of south-west Germany and has intensive agricultural activity although, as it happens commonly in Germany, it is densely populated. A set of 32 sub-basin, displayed in figure 1, were chosen to perform the analysis described in the following sections.

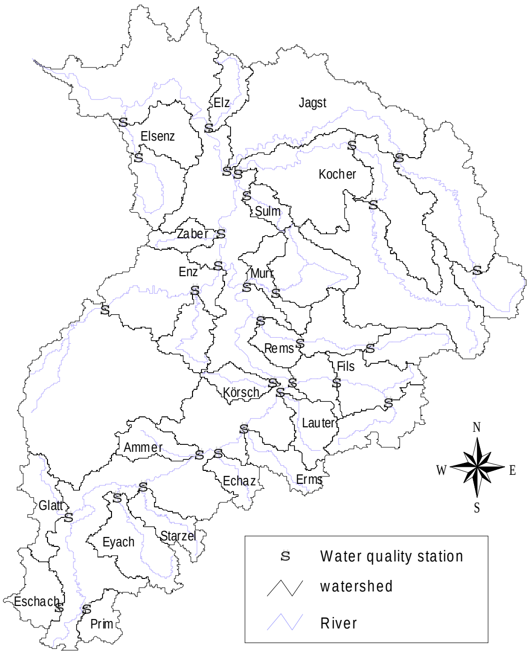


Figure 1.- The Neckar river basin and sub-basins

2.3 Database

The basis for the database consists of multispectral imagery, a digital terrain model (DTM) of the basin, soil maps, statistical data and hydrological monitoring data (see table 1). This information was stored and processed in a GIS in form of information layers. Since the basin comprises an extent area, the database is very big. GIS functions were used to estimate spatially referenced catchment attributes. The availability of the information in form of GIS layers enabled its combination via mathematical operations of layers or overlay. Table 1 shows a summary of the information sources and the computed parameters, that are discussed in more detail in the following paragraphs.

Table 1.- Database

Data Source	Information
Multispectral satellite imagery	Soil Use
	Roughness Coefficient
Digital terrain model	Slope
	Flow direction
	Basin Limits
	River Network
	Topographic Index
	Flow Time
	Distance to River Network
Statistical Data	Population
	Livestock
	Industry Type
	Agricultural Production
Soil Map	Infiltration Capacity
	Erodibility
Hydrological Monitoring	Water Quality Data
	Rainfall Data
	Flow Data

Multispectral imagery:

A key point in water quality is associated to human activities developed in the basin. Man is able to introduce great changes in the landscape and alter hydrological and chemical properties of the basin. Since the survey of pollution sources and land use is very expensive and time consuming, information about land cover and soil use was obtained classifying multispectral satellite imagery. The spectral data set comprises Landsat imagery captured in 1993. Soil use within the basin was classified in 16 classes. The resulting image, with spatial resolution of 30 meters, was introduced into the GIS environment for its combination with the other digital data.

Topography:

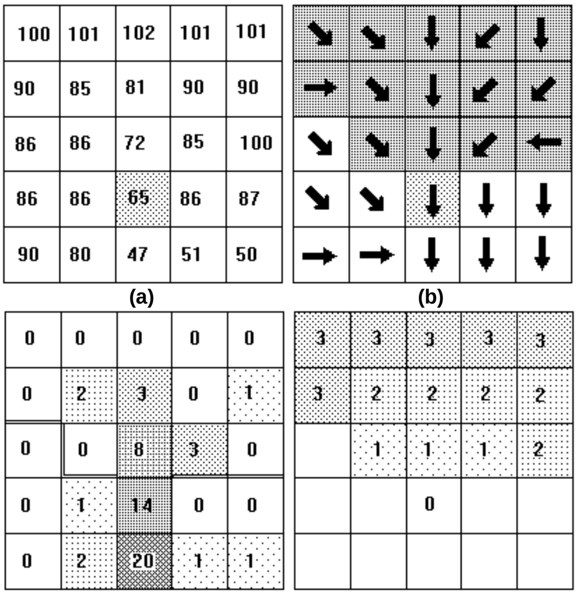
DTM are commonly used for the computation of topographical information as local slope, aspect or altitude profiles and have gained recognition in the field of spatial modeling. They can be interpolated from contour line maps or obtained using laser scanner technology. In the study presented here was used a 30 meters resolution grid, compatible with the resolution of the satellite imagery. New information layers were computed from the DTM: slope and flow direction, drainage network, specific catchment and distance to the drainage network and to the outlet.

In order to compute some of the parameters, flow was simulated over the DTM. The flow direction is based on the algorithm described by Jenson and Domingue (1988). The method considers that each pixel discharges into one of its eight neighbours, the one located in the direction of steepest gradient (figures 2a and 2b). The local gradient determines the flow direction at each pixel and the set of directions over the matrix defines flow paths (figure 2.b), which have the constraint that the flow occurs only downwards, from an upper cell to a lower one or to a cell with the same elevation. The obtained paths must also end at the borders of the elevation grid. Since a DTM may have depressions where an uphill flow would be necessary, it is practical to identify and “fill” them before estimating the flow paths. This task is performed within an iterative process, marking pixels with undefined flow direction and filling their watershed up to a value that satisfies the imposed restrictions.

The catchment of a selected point within the raster grid can be obtained grouping all pixels whose flow reaches this point (in this study, this point is generally the location of the water quality sample station). This is done following the flow paths upwards, starting at the outlet, till a cell is reached, which does not have any neighbours discharging into it (hatched area in figure 2b).

Information about the drainage is obtained from the specific drainage area of each pixel (figure 2c). Since the flow paths converge to the outlet, the pixels that receive more discharges are associated to the stream channels. A synthetic drainage network can be estimated from the flow accumulation grid.

The distance to the outlet along the flow paths can be also computed, following the flow directions along the matrix. The automatic extraction of catchment properties from DTMs has been emphasised by Jenson and Domingue (1988) and Donker (1992).



100	101	102	101	101
90	85	81	90	90
86	86	72	85	100
86	86	65	86	87
90	80	47	51	50

(a)

↘	↘	↓	↙	↓
→	↘	↓	↙	↙
↘	↘	↓	↙	←
↘	↘	↓	↓	↓
→	→	↓	↓	↓

(b)

0	0	0	0	0
0	2	3	0	1
0	0	8	3	0
0	1	14	0	0
0	2	20	1	1

(c)

3	3	3	3	3
3	2	2	2	2
	1	1	1	2
		0		

(d)

Figure 2.- Information derived from a DTM. (a) DTM; (b) flow paths; (c) specific drainage area; (d) distance to the outlet

Statistical data:

Another part of the database is related to statistical information. The Statistical Office of Baden-Württemberg provides wide information about the activities in the basin. Variables considered relevant for water quality were chosen: Population, livestock production, industry type, industrial production and agricultural production.

Hydrological data:

Data from the rainfall gauges were also utilised to interpolate a grid, which was integrated to the raster data set. Data for the period 1993-1996 were supplied by the national meteorological institute (Deutsche Wetterdienst - DWD).

Water quality and flow measurements within the same period were obtained from the State Office for Environmental Protection LfU (Landesamt für Umweltschutz from Baden-Württemberg/Germany).

As an example, two layers of the Elsenz sub-basin are shown: figure 3 displays the land use layer and figure 4 the livestock layers.

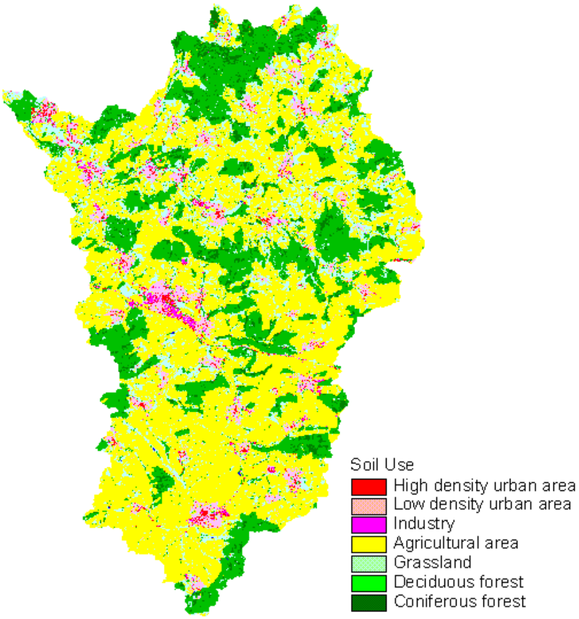


Figure 3.- Soil use in Elsenz sub-basin

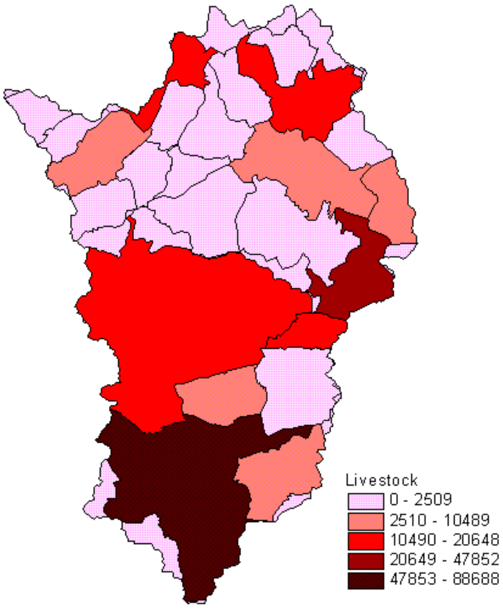


Figure 4.- Livestock distribution in Elsenz sub-basin

2.4 GMDH Algorithm

We used the GMDH (Group Method of Handling Data) of Ivakhnenko, which original version is described in Farlow (1984), to model the relationship between nitrate and the variables in the database. The model is self-organizing and is based on the statistical learning networks approach. It selects the set of variables automatically that better represents the dependent variable through a set of observations in an iterative process. Different data can be used as input in form of a vector of independent variables x, with n observations. Since the structure of the model is based on the input data set, the choice of this set is a very important step. The GMDH algorithm, either in its original form or one of its variations, has been used to solve problems in some areas as water quality, air pollution, hydrology, meteorology and economy. Examples of application in water quality are given in Müller (1996), Müller (1994), Tamura & Halfon (1980) and Duffy & Franklin (1975).

The model assumes that the independent variable y can be described by a polynomial combination of the m variables x. Since the polynomial is unknown, the algorithm starts computing the regression equation for the simple relation:

$$y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_i x_j$$

for each pair of variable (x_i, x_j). Thus, a new set of m(m-1)/2 variables is generated. The estimates of the independent variable is then compared to the observed values and the combinations with higher correlation values are kept, the rest is discarded. The selected variables are then used to replace the original ones and the process is repeated until the error, given by the difference between the observed and computed values, reaches a minimum, which characterises the best polynomial for the model. The philosophy of the algorithm is to selected the best combinations of variables at each iteration and use them to obtain a higher order polynomial for the next iteration.

Substituting the computed values it is possible to write the polynomial in terms of the original variables x. The polynomial of Ivakhnenko is in the form:

$$y = a + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} d_{ijk} x_i x_j x_k + \ldots$$

where:

a, b_i, $c_{i,j}$, d_{ijk} ...	: parameters of the polynomial
m	: number of independent variables
x_i, x_j, x_k, ...	: independent variables
y	: dependent variable

3 RESULTS

The information stored in the database was used to predict the nitrate concentration with the GMDH model. It must be pointed out that the input variables must be carefully selected, since they should have influence on the modeled parameter in order to obtain a reasonable result. For nitrate the following data were selected: the percentage of each land use class, distance from each soil use to the drainage network and rainfall. An input vector was computed for each sub-basin, which describes its characteristics.

32 water quality stations distributed all over the Neckar river basin were selected and the above mentioned parameters were computed. As a result, a vector for each station was obtained, which stores the spatial index, derived from the satellite imagery and DTM, and statistical information. This vector constitutes the input for the water quality model used to simulate nitrate concentration. Table 2 shows the summary of the input data.

Table 2.- Data for 32 study sub-basins.

Variable	Minimum	Maximum
Area (km²)	102	2213
% high density urban area	0	7
% low density urban area	2	20
% industry area	0	3
% agricultural area	13	61
% Grassland	10	39
% Deciduous forest	7	45
% Coniferous forest	2	45
Rainfall (mm annual)	773	1369
Distance of high density urban area to drainage (m)	390	870
Distance of low density urban area to drainage (m)	390	900
Distance of industry area to drainage (m)	450	1170
Distance of agricultural area to drainage (m)	660	1020
Distance of Grassland to drainage (m)	630	1200
Distance of Deciduous forest to drainage (m)	630	870
Distance of Coniferous forest to drainage (m)	690	1170
Population	27496	962952
Livestock	5636	717997
Nitrate Concentration (mg/l)	3,67	16,44

The GMDH identified the most significant variables that influence annually nitrate concentration:

- percentage of agricultural area and
- percentage of low density urban area

Figure 5 displays the comparison between the model output and the observed values for 1993. A correlation factor of 0.9 was found. These two most significant variables identified by GMDH algorithm are, too, in accordance with the commonly described sources of nitrate. Moreover, as reported also in Beudert (1997), the influence of rainfall was recognized as insignificant.

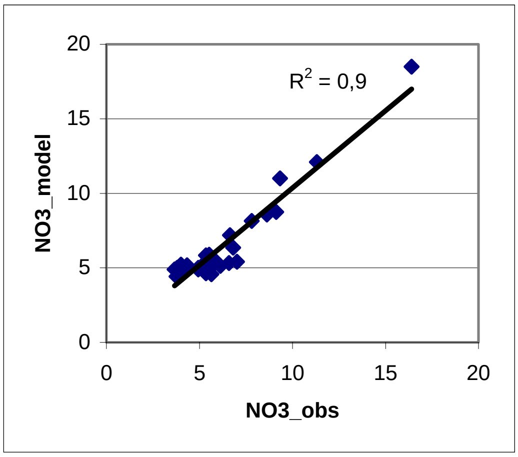


Figure 5.- Observed and predicted nitrate concentration

4 DISCUSSION

In this study we point out the importance of the use of spatial distributed data in water quality analysis using GIS together with a GMDH model. GIS proved to be a very efficient tool to resume spatial data and generate inputs for the model. The use of GIS for spatial analysis in water quality is also reinforced by the fact that the amount of available digital data is increasing. GIS speeds this process and enables a more accurate description of the basin.

Experiments made with GMDH and the computed data showed that the model helps to identify the most significant variables. The results enables a description of water quality in terms of easy available spatial and statistical data.

The spatial resolution of the raster data deserves special attention, since it has strong influence on some parameters, as slope, which may change according to variations on the resolution.

Because of the nature of the model a high number of iterations would provide a high order polynomial that best fits the observed values. Nevertheless this solution may not be optimal, because it is able to describe with high accuracy just the input data set and not the general trend of the basin.

The methodology was used to model nitrate, but new simulations will be performed for other water quality parameters, such as total phosphorus or heavy metals. These parameters depend on sediment transport, therefore, other model inputs should be added, according to the involved processes, as erosion, runoff and sediment tansport.

Acknowledgements: The authors would like to thank for the support of CNPq/Brazil.

5 REFERENCES

Beudert, G.. 1997. Gewässerbelastung und Stoffaustrag von befestigen Flächen in einem kleinen ländlichen Einzugsgebiet. Dissertation, Universität Karlsruhe, Oldenbourg Verlag, München. 216 p.

Cressie, N. and J.J. Majure. Spatio-temporal statistical modeling of livestock waste in streams. J. Agric. Biol. Environ. Stat., 2, 24-47, 1997.

Dillon, P.J. & L.A. Molot. Effect of landscape form on export of dissolved organic carbon, iron, and phosphorus from forested stream catchments. Water Resour. Res., 33, 11, pp. 2591-2600, 1990.

Donker, N.H.W. Automatic extraction of catchment hydrologic properties from digital elevation data. ITC Journal, pp. 257-265, 1992-3.

Duffy, J.J. and M.A. Franklin. A learning identification algorithm and its application to an environmental system. IEEE Trans. Syst., Man, Cybern., vol. SMC-5, N.2, pp. 226-240, March 1975.

Farlow, S.J.. 1984. Self-Organizing Methods in Modeling. New York and Basel: Marcel Dekker, Inc., 350p.

Jenson, S.K. and J.O. Domingue. Extracting topographic structure from digital elevation data for geographic information system analysis. Photogram. Eng. and Rem. Sens. Vol. 54, No. 11, pp. 1593-1600, 1988.

Jordan, T.E., D.L. Correll and D.E. Weller. Relating nutrient discharges from watersheds to land use and streamflow variability. Water Resour. Res., 33, 11, pp. 2579-2590, 1990.

Müller, J.-A.. Analysis and prediction of ecological systems. SAMS, Vol. 25, pp. 209-243, 1996.

Müller, N.. 1994. Gewässergütemodellierung von Fließgewässern unter Berücksichtigung qualitativer, quantitativer, flächenhafter und sozioökonomischer Informationen. Karlsruhe: Institut für Siedlungswasserwirtschaft, 155p.

Tamura, H. & E. Halfon. Identification of a dynamic lake model by the group method of data handling: an application to lake Ontario. Ecol. Modelling, 11:81-100, 1980.

Wolock, D.M.; G.M. Hornberger & T.J. Musgrove. 1990. Topographic effects on flow path and surface water chemistry of the Llyn Brianne catchments in Wales. Journal of Hydrology, 115, pp. 243-259.