UNIVERSITY OF SZEGED
Faculty of Sciences
Doctoral School of Earth Sciences
Department of Physical Geography and Geoinformatics

Artificial neural networks and geographic information systems for inland excess water classification

Theses of PhD Dissertation

Boudewijn van Leeuwen

Supervisors:

Dr. József Szatmári
Assistant professor

Prof. Dr. Gábor Mezősi
Professor

Szeged, 2012

Introduction

Due to its geographic position and climate, the Great Hungarian Plain is under continuous threat of droughts and floods. The year 2010 was one of the wettest years ever in Hungary. In the period October 2009 – December 2010, on the Great Hungarian Plain, 1149 mm precipitation fell, which corresponds to a yearly precipitation of 919 mm, while the long term average yearly precipitation is 489 mm (in Szeged). The extreme precipitation caused exceptionally large areas to be flooded by *inland excess water*. The maximum total flooded area during this 15 months long period was 355 000 ha on December 9, 2010 and the estimated financial damage to the agricultural sector alone exceeded 500 million Euros. Together with the consequential damage like soil degradation, inland excess water is one of the most severe natural hazards in the Carpathian basin. To be able to prevent or reduce damage due to inland excess water it is necessary to understand why and where it occurs.

There is no formal or official English word to describe the hydrological process that is the main topic of the dissertation. *Inland excess water* is a translation of the Hungarian word *belvíz* and will be used throughout this work. Although inland excess water got most scientific attention in Hungary, the phenomenon is not limited to this geographic region. For example, in China, India, Italy, Germany, the Netherlands, Serbia, Romania, and Russia it is occurs as well. The large number of definitions of inland excess water used in literature reflects the many scientific fields that deal with inland excess water research. Every field e.g. water management, agriculture, ecology, landscape planning or economics defines the

phenomenon from its own perspective. My general working definition that is used throughout this work is the following:

> *Inland excess water is water that temporarily remains in local depressions because of a combination of a surplus of water due to lack of runoff, insufficient evaporation and low infiltration capacity of the soil or because of upwelling of groundwater.*

Different genetic types of inland excess water can be distinguished: (1) The vertical type, which is caused by the upwards push of groundwater, (2) the horizontal type, that occurs due to precipitation and/or melting water that accumulates in local depressions because there is insufficient runoff, evaporation and/or infiltration, and (3) the type that occurs due to inland excess water that is transported from other areas towards a main river, but queues up in front of a pumping station because the station does not have enough pumping capacity.

Inland excess water is caused by a multitude of interrelated factors. They can be split into two groups; static factors that are stable over a period of decades or longer, such as relief and soil, and dynamic factors, which change within hours or days, like meteorology and groundwater level. Relief influences the runoff: water may collect in local depressions. Soil characteristics determine the infiltration and storage capacity. The amount of precipitation, as source of inland excess water is part of the meteorological factor. Other meteorological components like temperature and water vapour influence the evaporation rate. High groundwater levels may cause floodings, but can also prevent water from infiltration in the soil. Often anthropogenic factors strongly affect the formation of inland excess water. This can be due to the obstruction of runoff (e.g. by buildings, roads

or levees) and by reducing infiltration (e.g. due to paved surfaces), or by decreasing the chance of inundations due to for example the construction of channels and reservoirs.

Damages caused by inland excess water vary in time and space. In some years, over 10% of the agricultural land is flooded, while in other years there is hardly any damage. Some areas suffer from inland excess water in one year, while in other years, at the same place severe drought occurs. This complicates the possibilities to find solutions for the problem. The spatial pattern of inland excess water is also heterogeneous. Due to the spatial variation of the interrelated factors, areas that are never inundated can be found close to areas that are often under water.

To analyse the complex problem of inland excess water, it is important to understand its spatial and temporal distribution. This is done in two ways: (1) By mapping the spatial and temporal distribution based on in situ or remotely sensed observations, and (2) by estimating the impact of a selection of principle factors and weighing those using experimentally derived coefficients. Most studies have tried to identify the above factors and combined them using regression functions or other linear statistical methods. These methods have the disadvantage that they cannot deal with the nonlinear and complex functional relationships between those factors. Here, we present a different approach to identify and forecast inland excess water inundations using artificial neural networks (ANN) combined with geographic information systems (GIS). This approach has many advantages. First, it is independent of the statistical distribution of the data and there is no need to define the weight of the individual factors. Neural networks allow the target classes to be defined in relation to their distribution in the corresponding domain of each data source, and therefore the integration of

remote sensing and GIS data is very convenient. Furthermore, ANNs are capable of incorporating uncertainty, incomplete data, incorrect sampling, multicollinearity between variables, spatial or temporal autocorrelation, and insignificance of single variables. These are common in geographic analysis, but especially in inland excess water research due to the fuzzy nature of the boundaries of the inundations and the complex interrelations between the factors.

Data and Methods

To facilitate the efficient application of classification of inland excess water occurrences by artificial neural networks, an integrated GIS – ANN framework was created using a combination of *ArcGIS*, a geographic information system, *Matlab*, a mathematical modelling software and *Python*, an open source programming language (Fig 1).

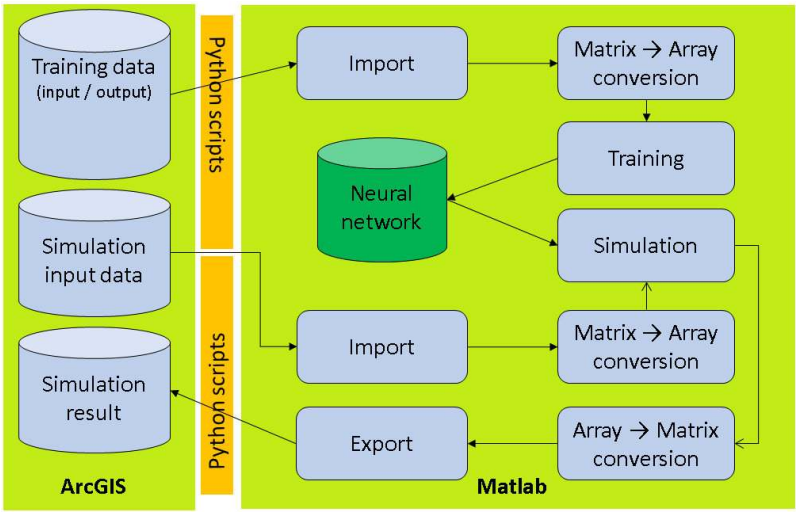

Figure 1. ANN – GIS framework showing the workflow in ArcGIS and Matlab

The framework was created to handle input data, intermediate results and output data in a flexible way in both ArcGIS and Matlab. In this way, it is possible to create the data files, test different network settings, perform training and simulation, and evaluate and visualize the training and simulation results efficiently. All steps are executed from within the GIS and no direct user interaction with the ANN software is needed.

The framework was developed and evaluated with data from a 20 km^2 area near Szeged, Hungary. The area is suitable for inland excess water research for two main reasons: (1) the soils in the area show extreme mechanical properties. The bad permeability characteristics combined with the very flat terrain with large local depressions without runoff, result in high vulnerability to inland excess water accumulation and (2) the Department of Physical Geography and Geoinformatics has a long standing inland excess water research program in the area. This means that lots of data and knowledge about the area are available. Furthermore, the area is close to the airport of Szeged from where the data acquisition campaigns are executed. Since inland excess water is a phenomenon with strong temporal characteristics, it is important to acquire data for research in time.

Five input data sets were created to be used in the framework: (1) Colour infrared (CIR) images collected with an in-house developed acquisition system based on a MS3100 digital camera, (2) a 1 m resolution LIDAR based digital elevation model, (3) a 1:25 000 scale soil database, (4) a database of anthropogenic objects in the area (channels, roads, buildings and oil wells), and (5) inland excess water ground truth data collected with hand-held GPS systems. Other input data that may influence the formation of inland excess water were not incorporated for several reasons. First, soil

measurement showed everywhere in the area poor permeability characteristics, and differences in other soil characteristics were also minimal. Therefore, the soil is considered homogeneous in the study areas. The lithology is also considered to be homogeneous throughout the area. A groundwater-precipitation-evaporation measuring station in the study area provides hourly data of many parameters. These parameters show that the precipitation, evaporation and infiltration result in a surplus of water during the inland excess water periods. Vegetation is not homogeneously distributed over the area, but was not introduced as a separate input layer to the network, because its distribution is represented by the colour infrared images.

Results and conclusions

1. The dissertation presents the first extended theoretical description of inland excess water in English. This is an important step in the creation of awareness of the problem beyond the Hungarian language region. In many bordering countries the same phenomenon occurs, but there, e.g. in Serbia and Romania, the problem is treated similarly to river floodings and no special attention is given to its underlying physical processes.
2. To evaluate the usefulness of in situ observations by the regional water directorates as accurate ground truth for the neural network classification, they were statistically compared with local depressions calculated from a LIDAR derived digital elevation model. The base hypothesis was that there exists a positive relation between the number of occurrences of inland excess water and the depths of the local depressions. This relationship between the two data sets was not found.

Therefore, it was concluded that the in situ observation map could not be used as base data for the framework.

3. An image processing workflow has been developed to convert raw LIDAR data and digital aerial photographs into useful standardized input data for the artificial neural network classification.
4. Individual inland excess water patches were measured in the field with different types of GPS systems. Only a small difference (1,8 %) in the area of the individual inundations was found between the measurements from the used instruments. The area differences caused by the differences in accuracy of the GPS systems are smaller than the inaccuracy caused by the data collection method. Therefore, it can be concluded that for the measurements of individual patches, the less accurate (and cheaper) GPS systems are sufficient.
5. For the first time, an approach is presented and successfully carried out, that identifies and (depending on the opportunities) predicts inland excess water inundations using artificial neural networks combined with geographic information systems.
6. It is the first time that a GIS – ANN framework is created using a combination of ArcGIS (GIS), Matlab (modelling software), and Python (programming language) that is fully integrated, which means that all interactions with the ANN software are executed from within the GIS.
7. Due to the nature of artificial neural networks and their training algorithms, the method can be calculation intensive. Furthermore, in inland excess water analysis, the spatial data that is used consists of many high resolution layers. The combination of large data sets and heavy calculation demand results in long training times and memory

problems. To overcome these problems, a data reduction algorithm has been developed that reduces the amount of input data by taking only every n^{th} pixel of the input data layers during the conversion of the layers into the matrix that is fed to the neural network. The use of the data reduction factor has also a positive side effect on the training quality. Due to the reduced amount of input samples, over-fitting by the network is avoided.

8. The ANN – GIS framework is a calculation intensive approach. Experiments have been executed to determine possibilities for performance improvements. It was found that there is an exponential relationship between the time needed for training of the network and the amount of data in the training data set. It was also found that there is an exponential relationship between the time needed for training and the amount of data that is written to the computer's memory (Fig 2). Based on these observations, the Data reductions factor was developed and the Memory reduction function (of Matlab) was used.

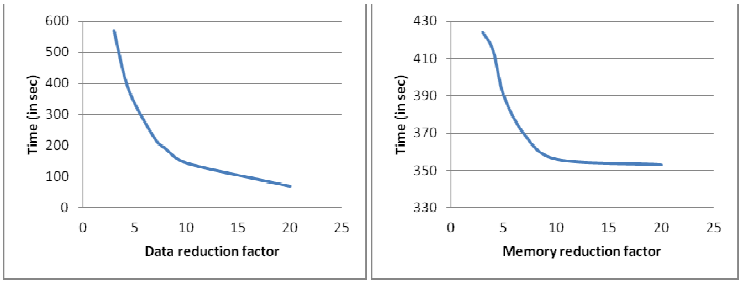

Figure 2. The relation between the time needed for 50 iterations during the training phase and the data reduction factor, and the training time and the memory reduction function (on the standard analysis computer with an Intel Core 2, 2.6 GHz processor and 2 Gb RAM)

9. Four simulations were executed to evaluate the influence of the different input layers. Each simulation was executed with a different set of input layers, but with the same neural network settings. The input layers are given in table 1, and the output results are shown in figure 3.

Table 1. Input data for training and simulations on the training area

Description			
1	Local depressions	6	Distance from buildings
2	Agrotopo soil characteristics	7	Aerial photograph band 1
3	Distance from channels	8	Aerial photograph band 2
4	Distance from roads	9	Aerial photograph band 3
5	Distance from oil wells		

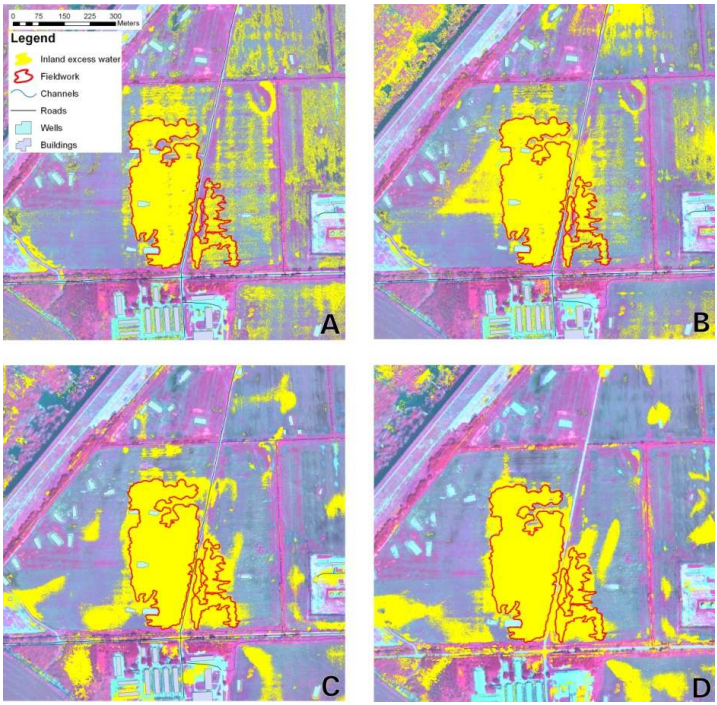

Figure 3. Evaluation of the input layers. All results cover the same area at the same scale

Table 2. Spatial correlation between input layers, training layer and simulations

	A	B	C	D	Train	1	2	3	4	5	6	7	8	9
A	1	0,91	0,83	0,83	0,76	-0,79	0,07	-0,3	0,14	0,13	0,06	0,23	0,02	0,04
B	0,91	1	0,91	0,91	0,83	-0,73	0,06	-0,44	0,11	0,14	0,03	0,22	0,02	0,04
C	0,83	0,91	1	0,99	0,92[b]	-0,66	0,09	-0,4	0,18	0,17	0,11	0,19	0,02	0,03
D	0,83	0,91	0,99	1	0,92[c]	-0,66	0,09	-0,4	0,18	0,17	0,11	0,19	0,02	0,03
Train	0,76	0,83	0,92	0,92	1	-0,59	0,08	-0,37	0,16	0,16	0,09	0,18	0,02	0,03
1	-0,79	-0,73	-0,66	-0,66	-0,59[a]	1	-0,05	0,19	-0,1	-0,06	-0,05	-0,12	-0,09	-0,12
2	0,07	0,06	0,09	0,09	0,08[d]	-0,05	1	0,28	0,46	0,33	0,34	0,03	-0,06	-0,06
3	-0,3	-0,44	-0,4	-0,4	-0,37	0,19	0,28	1	0,27	0,07	0,31	-0,23	0,01	-0,02
4	0,14	0,11	0,18	0,18	0,16	-0,1	0,46	0,27	1	0,14	0,71	0,03	-0,07	-0,05
5	0,13	0,14	0,17	0,17	0,16	-0,06	0,33	0,07	0,14	1	-0,03	0,05	-0,01	-0,04
6	0,06	0,03	0,11	0,11	0,09	-0,05	0,34	0,31	0,71	-0,03	1	0,04	-0,11	-0,05
7	0,23	0,22	0,19	0,19	0,18	-0,12	0,03	-0,23	0,03	0,05	0,04	1	-0,13	0,26
8	0,02	0,02	0,02	0,02	0,02	-0,09	-0,06	0,01	-0,07	-0,01	-0,11	-0,13	1	0,63
9	0,04	0,04	0,03	0,03	0,03	-0,12	-0,06	-0,02	-0,05	-0,04	-0,05	0,26	0,63	1

The simulations have been performed on data from the training area only, since only there, it is possible to compare the results with ground truth data. The first simulation (A) is only based on the CIR images and the local depressions. The second (B) also incorporates the distances to channels, the third simulation (C) incorporates 8 input layers, only soil was excluded. The final simulation (D) includes all 9 input layers. The spatial correlations between the different input layers and output results show the importance of the different layers in the simulation (Table 2). The first simulation result clearly shows the depressions in the area (Fig. 3A). These depressions have a correlation of -0,59 with the fieldwork or training data (Table 2. a). This correlation is negative because the depression classes range from *no depression* to *deep depression* and the inland excess water ranges from *no inland excess water* to *inland excess water*. The relatively high value shows that the relief has a strong influence on the formation of inland excess water. In general, it can be concluded that the more layers are added to the simulation the better the result. The best result can be seen at simulations C and D. The only difference between these simulations is the soil layer. Adding this layer to the simulation slightly reduces the spatial correlation (Table 2. b and c). The limited influence of the soil in the simulations in this study area is also reflected in the low spatial correlation (0,08) between the fieldwork data and the soil map (Table 2. d).

10. From the overall accuracy and Cohen's Kappa calculations, it can be concluded that in general, more layers in the training and simulation phase result in an increase of the overall quality of the simulations

(Table 3), although when the soil layer is added, this has a negative effect on the quality (see 9th point).

Table 3. Cohen's Kappa and the overall accuracy of 4 simulations

	A (4 layers)	B (5 layers)	C (8 layers)	D (9 layers)
Cohen's Kappa (κ)	0,76	0,81	0,86	0,83
Overall accuracy	0,88	0,91	0,93	0,91

11. The ANN classifications were compared with traditional minimum distance (MD) classification and maximum likelihood (ML) classifications (Table 4). The ANN – GIS framework outperforms the traditional classifications, even in the case when only 3 layers are used.

Table 4. Overall accuracies of different types of classifications

	Overall accuracy
MD based on 2 classes	67 %
ML based on 2 classes	69 %
ML with merged non water classes	70 %
ANN with two classes (3 layers)	74 %
ANN with two classes (8 layers)	93 %

12. The result of the maximum likelihood classification and the ANN with only three remote sensing bands is quite similar (Table 4). Improving this result can be achieved by adding extra information, like local depressions, distance to anthropogenic objects and soil type to the classifications. Adding these extra input layers is only possible with the ANN approach since the additional layers are not compatible with the remote sensing data, and therefore it is not possible to improve the maximum likelihood method in this way.

Own publications relevant for the theses

Van Leeuwen B., Mezősi G., Tobak Z., Szatmári J., Barta K., 2012, Identification of inland excess water floodings using an artificial neural network, Carpathian Journal of Earth and Environmental sciences. (In review).

Van Leeuwen B., Tobak Z., Szatmári J., 2008, Development of an integrated ANN - GIS framework for inland excess water monitoring, Journal of Env. Geogr. I, 3-4, pp. 1-6.

van Leeuwen B., Tobak Z., Szatmári J., Barta K., 2010, Új módszerek alkalmazása a belvizek keletkezésének vizsgálatában és monitorozásában, In: Lóki J., Demeter G. (Eds.) Az elmélet és gyakorlat találkozása a térinformatikában I, Debrecen. pp. 121-130.

van Leeuwen B, Tobak Z, Szatmári J, Mucsi L, Fiala K, Mezősi G, 2009, Small format aerial photography: a cost effective approach for visible, near infrared and thermal digital imaging, In: Car A, Griesebner G, Strobl J (Eds.), Geospatial crossroads, Heidelberg: Herbert Wichmann Verlag, pp. 200-209.

Szatmári J., Szíjj N., Mucsi L., Tobak Z., van Leeuwen B., Lévai Cs., Dolleschall J., 2012, Comparing LIDAR DTM with DEM-5 of Hungary, In: Geiger J., Pál Molnár E., Malvic T. (Eds.), New horizons in Central European geomathematics, geostatistics and geoinformatics, Institute of Geosciences, University of Szeged, pp. 151-158.

Szatmári J., Szíjj N., Mucsi L., Tobak Z., van Leeuwen B., Lévai Cs., Dolleschall J., 2011., A belvízelöntések térképezését és a belvízképződés modellezését megalapozó térbeli adatgyűjtés, In: Lóki J. (Ed.), Az elmélet és gyakorlat találkozása a térinformatikában II., Debrecen. pp. 27-35.

Szatmári J., Tobak Z., van Leeuwen B., Dolleschall J., 2011, A belvízelöntések térképezését megalapozó adatgyűjtés és a belvízképződés modellezése neurális hálózattal, Földrajzi Közlemények 135, 4, pp. 351–363.

Tobak Z., Szatmári J., van Leeuwen B., 2008b, Small format aerial photography – remote sensing data acquisition for environmental analysis, Journal of Env. Geogr I, 3-4, pp. 21-26.

Unger J., Gál T., Rakonczai J., Mucsi L., Szatmári J., Tobak Z., van Leeuwen B., Fiala K., 2010, Modeling of the urban heat island pattern based on the relationship between surface and air temperatures, Időjárás 114, 4, pp. 287-302.