STUDIA INFORMATICA
Nr 1/2(7) Systemy i technologie informacyjne 2006

Estimation, Decoding and Forecasting in HMM and Hybrid HMM/ANN Models: a Case of Seismic Events in Poland

Katarzyna Bijak[1]
[1] Retail Credit Risk Department, Bank BPH SA
Emilii Plater Street 53, 00-113 Warsaw, Poland

Abstract. This paper compares performance of a hidden Markov model (HMM) and a hybrid HMM/ANN model in seismic events modeling. Observation variables are assumed to follow a Poisson distribution. Parameters of the discrete-time two-state models are estimated on the basis of data on seismic events that were recorded in Poland from 1991 to 1995. Then, on the basis of the estimation results, the most likely sequences of states of the hidden Markov chains are found and forecasts for January 1996 are made. It is shown that the hybrid model fits better to the data.

Keywords. Hidden Markov models, hybrid HMM/ANN models, neural networks, seismic events

1 Introduction

Hidden Markov models (HMMs) are models with unobserved (hidden) Markov chains. They were introduced in 1966 by Baum and Petrie [3]. Since then HMMs have been widely used in speech recognition and signal modeling [4, 9]. Nowadays it is believed that they are one of the best and most successful acoustic models [10]. According to [2], HMMs have been applied in time series analysis since the mid 1970s.

Hybrid HMM/ANN models combine HMMs and artificial neural networks (ANNs). The hybrid models emerged in the beginning of the 1990s and the goal was to take advantage from the features of both HMMs and ANNs. A variety of different architectures and training algorithms have been proposed for these models [13]. There are, among other architectures, hybrid models with perceptrons and multilayer perceptrons (HMM/MLP), radial basis function networks (HMM/RBF), self organizing maps (HMM/SOM), recurrent neural networks (HMM/RNN) as well as time delay neural networks (HMM/TDNN) [7]. Those neural networks play different roles in hybrid models: they provide distribution parameters, emulate HMMs, transform observations into a form that is more suitable for HMMs, perform quantization of signals in signal models and have some other functions [13].

So far their applications hardly go beyond speech recognition and signal modeling. There are, however, some papers on using the hybrid models in time series analysis. For example Rynkiewicz [11] presents a hybrid HMM/MLP model for time series prediction.

Application of the hybrid models in time series analysis seems to be promising as they could fit better to the data than HMMs. Unlike in HMMs, in such models the parameters of observation distributions are not assumed to be constant. In the hybrid models these parameters depend on the previous observations through ANNs. Thus an additional information on the previous observations is used what cannot be done directly in HMMs. In this paper the performance of a HMM and a hybrid HMM/ANN model is compared in seismic events modeling for Poland.

Modeling and forecasting earthquakes (and catastrophes in general) is an important issue as far as risk assessment is concerned. Usually, the short-term forecasts are based on phenomena which are thought to be associated with seismic events, while the long-term forecasts assume cyclical nature of earthquakes. One of the biggest problems in this field consists in the fact that it is usually difficult to find out what caused an earthquake. One can observe seismic events in a given region but it can be assumed that behind them there is an unobserved and changing seismic activity of the region that influences occurrence of those events. Therefore, models with unobserved Markov chains seem to be suitable for seismic events modeling. The states of a hidden Markov chain could be referred to as the states of different (e.g. lower and higher) seismic activity.

2 Models

2.1 Hidden Markov Model

Let $\{C_t\colon t \in \mathrm{N}\}$ be an irreducible homogeneous Markov chain on the state space $\{1, 2\}$ with transition probability matrix:

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} = \begin{bmatrix} 1-\gamma_1 & \gamma_1 \\ \gamma_2 & 1-\gamma_2 \end{bmatrix} \tag{1}$$

where for all states i and j and times t:

$$\gamma_{ij} = \mathrm{P}(C_t = j \mid C_{t-1} = i)\,. \tag{2}$$

As $\{C_t\}$ is finite-state and irreducible there is a unique stationary distribution:

$$\boldsymbol{\delta}' = \begin{bmatrix} \delta_2 & \delta_1 \end{bmatrix} = \begin{bmatrix} \dfrac{\gamma_2}{\gamma_1+\gamma_2} & \dfrac{\gamma_1}{\gamma_1+\gamma_2} \end{bmatrix}. \tag{3}$$

The stationary distribution is taken as the initial distribution. The Markov chain $\{C_t\}$ is unobserved.

Let $\{S_t: t \in \mathrm{N}\}$ be such a random process that discrete random variables $\{S_t: t = 1, \ldots, T\}$ are mutually independent given $\{C_t: t = 1, \ldots, T\}$ and if $C_t = i$ then S_t follows a Poisson distribution with a mean λ_i:

$$\mathrm{P}(S_t = s \mid C_t = i) = {}_t\pi_{si} = \pi_{si} = \frac{e^{-\lambda_i}\lambda_i^s}{s!}. \tag{4}$$

The probabilities ${}_t\pi_{si} = \pi_{si}$ do not depend on time.

On the contrary to the Markov chain, the process $\{S_t\}$ is observed and thus values of the variables S_t are called observations. This model definition is similar to the definition given in [8].

Given the model, the probability of the observation sequence i.e. the likelihood function can be written as in [8]:

$$L_T = P(S_1 = s_1, \ldots, S_T = s_T) = \sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\ldots\sum_{ir=1}^{2}[({}_1\pi_{s_1 i_1}\ {}_2\pi_{s_2 i_2}\ \cdots {}_r\pi_{s_r i_r})(\delta_{i_1}\gamma_{{}_1 i_2}\gamma_{i_2 i_3}\ldots\gamma_{i_{T-1} i_T})]. \tag{5}$$

Estimation of the model parameters can be performed by maximizing the likelihood function. Another important function is the probability of the state sequence conditioned on the observation sequence [8]:

$$\mathrm{P}(C_1 = i_1, \ldots, C_T = i_T \mid S_1 = s_1, \ldots, S_T = s_T) = \frac{({}_1\pi_{s_1 i_1}\ {}_2\pi_{s_2 i_2}\ldots{}_T\ \pi_{s_T i_T})(\delta_{i_1}\gamma_{i_1 i_2}\ldots\gamma_{i_{T-1} i_T})}{L_T}. \tag{6}$$

Maximization of this function makes it possible to find the most likely sequence of states of the hidden Markov chain, given the observation sequence and the estimated model.

As far as forecasts are concerned, there is a useful formula for the expected value of the next observation S_{T+1}, given the observation sequence [8]:

$$\mathrm{E}(S_{T+1} \mid S_1 = s_1, \ldots, S_T = s_T) = \frac{1}{L_T}\boldsymbol{\delta}'{}_1\boldsymbol{\lambda}(s_1)\boldsymbol{\Gamma}_2\boldsymbol{\lambda}(s_2)\ldots\boldsymbol{\Gamma}_T\boldsymbol{\lambda}(s_T)\boldsymbol{\Gamma}_{T+1}\mathbf{L1} \tag{7}$$

where ${}_t\boldsymbol{\lambda}(s_t) = \mathrm{diag}([{}_t\pi_{s_t 1} \quad {}_t\pi_{s_t 2}]')$ and ${}_{T+1}\mathbf{L} = \mathrm{diag}([\lambda_1 \quad \lambda_2]')$.

In case of a two-state Markov chain the first entrance time from state i to state j is equal to $1/\gamma_i$ and the first return time to state i is equal to $1/\delta_i$.

2.2 Hybrid HMM/ANN Model

Let $\{C_t: t \in \mathrm{N}\}$ be a hidden Markov chain as it was defined in the previous section and let $\{S_t: t \in \mathrm{N}\}$ be such a random process that if $C_t = i$ then S_t follows a Poisson distribution with a mean $\lambda_i(\mathbf{s}_t)$ where $\mathbf{s}_t = [s_{t-2}\ s_{t-1}]'$ is a vector of two observations prior to S_t and $\lambda_i(\mathbf{s}_t)$ is the ith output of a neural network in case $\mathbf{s}_t$ is its input. Hence, the conditional distribution of S_t can be written as:

$$\mathrm{P}(S_t = s \mid C_t = i) = {}_t\pi_{si} = \frac{e^{-\lambda_i(\mathbf{s}_t)}(\lambda_i(\mathbf{s}_t))^s}{s!}. \tag{8}$$

Unlike in the HMM, the probabilities ${}_t\pi_{si}$ depend on time.

The neural network is called a match network as its outputs replace parameters of conditional observation distributions [10]. It is assumed to be a perceptron with two units both in the input layer and in the output layer. The architecture of this neural network is shown in Figure 2.1.

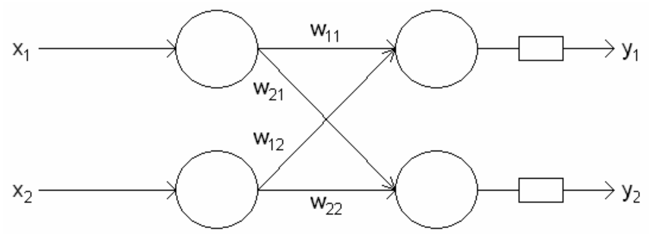

Figure 2.1. Architecture of a perceptron used in the hybrid model

In the input layer there are linear units that have an identity transfer function $y_k = f_{in}(x_k) = x_k$ where x_k is an input received by kth unit ($x_1 = s_{t-2}$ and $x_2 = s_{t-1}$). In the output layer units are nonlinear and the transfer function is the following:

$$y_i = f_{out}(x_i) = \begin{cases} \alpha \ln x_i + \alpha & \text{dla } x_i \geq 1 \\ \alpha e^{x_i - 1} & \text{dla } x_i < 1 \end{cases} \tag{9}$$

where α is a parameter such that $\alpha \in \mathrm{R}_+$ and x_i is a total input that the ith unit in this layer receives from the units in the previous layer:

$$x_i = \sum_{k=1}^{2} w_{ik} y_k + 1 \cdot \tag{10}$$

According to (10), in the output layer the threshold of a unit equals 1. Since the transfer function is such that $y_i \in \mathrm{R}_+$ and it is assumed that $\lambda_i(\mathbf{s}_t) = y_i$ for each state i of the Markov chain, the value $\lambda_i(\mathbf{s}_t)$ is real and positive and thus it can be a mean of a conditional Poisson distribution followed by S_t. Therefore, no additional constraints on neural network weights and distribution parameters are needed in the model estimation process. On the contrary to the hybrid models, the HMMs lack that advantage. As in case of the HMM, the process $\{S_t\}$ is observed.

Parameters of the hybrid model can be estimated altogether including weights of the neural network and the estimation can be performed using the maximum likelihood method [1, 12]. Both (5) and (6) are true for the hybrid model as well as for the HMM. The formula (7) is also valid for the hybrid model with the difference that ${}_{T+1}\mathbf{L} = \mathrm{diag}([\lambda_1(\mathbf{s}_T) \quad \lambda_2(\mathbf{s}_T)]')$.

Apart from match networks described in this section, there are transition networks that can also be introduced into a hybrid model. Outputs of a transition network replace transition probabilities and thus those probabilities are not constant [10]. Therefore, a hidden Markov chain is non-homogeneous in the hybrid model with a transition network.

Instead of one match or transition network, there can be a separate network for each state of the hidden Markov chain. Such networks have the same architecture for all states while their weights vary from state to state [10].

In between the HMMs and the hybrid models there are other modifications of the HMMs, namely models in which ${}_t\pi_{si}$ probabilities or transition probabilities depend on some additional variables such as, for instance, previous observations [8]. Modeling those dependencies using neural networks leads to development of the hybrid models (with match networks in the former case and with transition networks in the latter case).

3 Data

Seismic events in Poland have been recorded by the network of seismic stations for many years. The magnitude of the recorded events does not exceed 6 on the Richter scale. Most of these events are not tectonic earthquakes but collapse ones i.e. earthquakes in underground caves and mines. Seismic events occur mainly in southern and south-western Poland.

The data used in this research come from the Prototype International Data Centre and were downloaded from the web page of dr. Zbigniew Zwoliński from the Institute of Quaternary Research and Geoecology at Adam Mickiewicz University in Poznań [14]. The data concern number of seismic events that were recorded in Poland in successive months from January 1991 to December 1995 (60 observations). Additionally, the data for January 1996 were used in order to compare them ex-post with the forecast results. Only events with magnitude 3 or more are taken into consideration because they can be treated as rare phenomena and modeled using a Poisson distribution. Sample statistics, that were calculated on the basis of data used for the models estimation, are presented in Table 3.1.

Table 3.1. Sample statistics

Statistic	Value
Minimum	1
Maximum	15
Mean	6.897
Variance (variation coefficient)	9.743 (45.3%)

Monthly numbers of seismic events with magnitude 3 or more fluctuated between 1 and 15. In the period under study there were on average 6.897 such events a month. From the fact, that the sample mean does not equal the sample variance, it could be concluded that there is overdispersion in the data [5]. However, in case of the HMMs with Poisson distribution, overdispersion is an acceptable phenomenon as it can be shown that the mean of S_t does not have to be equal to its variance [8]. It seems that in case of the hybrid models overdispersion is not a problem, either. Tolerance of overdispersion is a significant advantage of the HMMs as that phenomenon is often observed in data.

4 Results

4.1 Estimation Results

On the basis of 58 observations (starting from March 1991) both models described in Chapter 2 are estimated. The observations for January and February 1991 are used as a neural network input in the hybrid model for March 1991. In the hybrid model the parameter α of the transfer function is assumed to equal 1.275 in the output layer.

The model parameters are estimated using the maximum likelihood method and the likelihood function is maximized using genetic algorithms. The likelihood function value is greater in case of the hybrid model ($8.089 \cdot 10^{-60}$) than in case of the HMM ($4.680 \cdot 10^{-61}$). The sum of squares is calculated as a sum of squares of differences between the observations and the expected values of S_t and once again a better result is obtained for the hybrid model: 460.839 in comparison with 565.352 for the HMM. Greater likelihood function value and nearly 20% lower sum of squares prove that the hybrid model fits better to the data.

The estimates of model parameters are shown in Table 4.1. The estimates of the parameters γ_1 and γ_2 are similar in both models (ca 0.02 and ca 0.06, respectively). This results in similar estimates of the transition probability matrix and the stationary distribution as well as the first entrance and return times.

Table 4.1. Maximum likelihood estimates of the model parameters

Parameter	Hidden Markov model	Hybrid HMM/ANN model
γ_1	0.024	0.022
γ_2	0.066	0.059
λ_1	5.419	-
λ_2	9.372	-
w_{11}	-	6.826
w_{12}	-	-1.168
w_{21}	-	105.944
w_{22}	-	-20.659

The estimates of the stationary distribution elements and other model characteristics are shown in Table 4.2. In case of the hybrid model the estimates of the parameters λ_1 and λ_2 for the successive months are calculated as outputs of the neural network with the estimated weights. In both models the estimate of the expected value of λ_2 equals ca 9 and is more than one and a half times greater than the estimate of the expected value of λ_1 (ca 5.5). Therefore, the first state of the hidden Markov chain could be called the state of lower seismic activity and the second state – the state of higher seismic activity. In the HMM the estimates of the parameters λ_1 and λ_2 are by definition constant while in the hybrid model they are characterized by the relatively small variation (16% and 12%).

Table 4.2. Estimates of the model characteristics

Characteristic	Hidden Markov model	Hybrid HMM/ANN model
δ_1	0.731	0.723
δ_2	0.269	0.277
The first entrance time from state 1 to state 2	40.917	44.484
The first entrance time from state 2 to state 1	15.044	17.050
The first return time to state 1	1.368	1.383
The first return time to state 2	3.720	3.609
Expected value of λ_1	5.419	5.730
Expected value of λ_2	9.372	9.092
Variance of λ_1 (variation coefficient)	0 (0%)	0.852 (16.1%)
Variance of λ_2 (variation coefficient)	0 (0%)	1.220 (12.1%)
Expected value of S_t	6.482	6.661*
Variance of S_t (variation coefficient)	9.554 (47.7%)	8.935 (45.4%)*

From the estimates of the stationary distribution elements δ_1 and δ_2 it could be concluded that according to both models, after a long time the probability of a Markov chain being in the state of lower seismic activity equals ca 0.73 and is nearly three times greater than the probability of a Markov chain being in the state of higher seismic activity (ca 0.27). Thus it appears that in Poland the dominating state is the state of lower seismic activity.

The estimates of the first entrance times could be interpreted in the following way: if Poland is in the state of lower seismic activity then the country is expected to move to the state of higher seismic activity in more than three years (over 40 months) while if Poland is in the state of higher seismic activity then the country is expected to move to the state of lower seismic activity in more than one year (over 15 months).

As far as the estimates of the first return times are concerned, the interpretation could be the following: if in a given month Poland is in the state of lower seismic activity then next time such a month is expected to be in ca 1.4 months (so it is likely that simply it will be the next month) while if in a given month Poland is in the state of higher seismic activity then next time such a month is expected to be in ca 3.7 months.

From the long first entrance times and short return times it could be concluded that in general periods of both higher and lower seismic activity are rather long in Poland.

Unlike in the HMM, in the hybrid model the estimates of the expected value and variance of S_t are not constant and therefore only their average values can be compared. The average estimates of the expected value and variance of S_t (ca 6.5 and ca 9, respectively) are similar in both models and in the sample. As far as seismic events with magnitude 3 or more are concerned, in Poland one could expect ca 6.5 such events a month.

4.2 The Most Likely Sequences of States of the Hidden Markov Chain

For both the HMM and the hybrid model, the most likely sequences of states of the hidden Markov chain, given the observation sequence and the estimated parameters, are found through maximization of the function describing probability of the state sequence conditioned on the observation sequence. Finding such sequences is called decoding [6]. As in the model estimation process, the target function is maximized using genetic algorithms.

As the estimates are similar for both models, the most likely sequences of states of the hidden Markov chain are similar, too. The observations and the expected values of S_t given these sequences are shown in Figure 4.1 (for the HMM) and in Figure 4.2 (for the hybrid model). Comparing these figures, one can notice that the expected values of S_t are generally closer to the observations in case of the hybrid model than in case of the HMM. In fact the sum of squares (calculated as a sum of squares of differences between the observations and the expected values of

S_t given the most likely sequence of states) is ca 10% lower for the hybrid model (287.209) than for the HMM (323.868).

In both cases the most likely sequence is such that the Markov chain is in the second state only in ca 20 successive times starting from the 16th time. It could be interpreted in the following way: Poland was likely in the state of higher seismic activity from April 1992 to the end of 1993, the beginning of 1994. In the rest of the period under study Poland was likely in the state of lower seismic activity. As expected, periods of both higher and lower seismic activity were quite long.

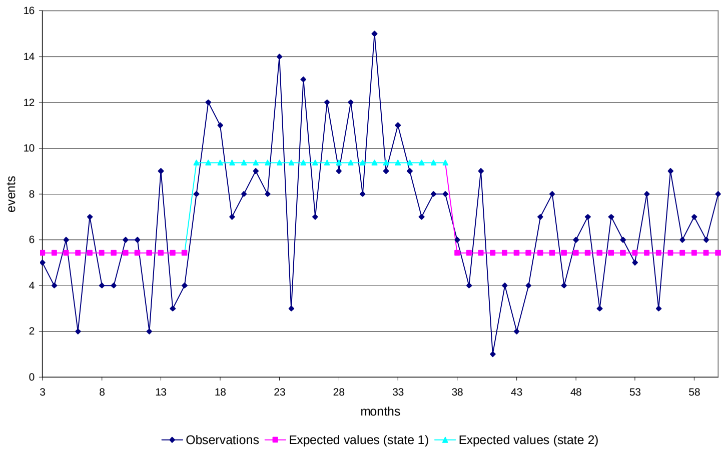

Figure 4.1. Observations and the expected values of S_t for the most likely sequence of states of the hidden Markov chain in the hidden Markov model

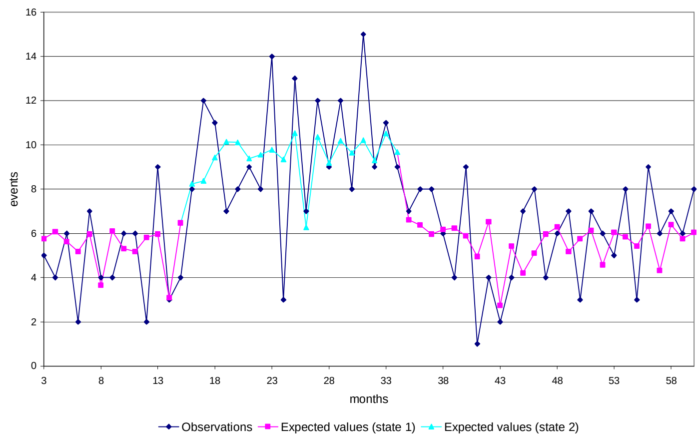

Figure 4.2. Observations and the expected values of S_t for the most likely sequence of states of the hidden Markov chain in the hybrid HMM/ANN model

4.3 Forecasts

Using estimates of the model parameters and the observation sequence makes it possible to forecast the value of next observation S_{T+1} (number of seismic events with magnitude 3 or more in Poland in January 1996). According to the HMM, the conditional expected value of S_{T+1} equals 5.862 while according to the hybrid model, it is equal to 6.028. In fact, in January 1996 there were 7 seismic events with magnitude 3 or more in Poland.

5 Conclusions

Using models with hidden Markov chains for seismic events modeling, one can distinguish states of different seismic activity of a given region. On the basis of data for Poland the states of lower and higher seismic activity are distinguished. Similarity between the estimates for the HMM and the estimates for the hybrid model suggests that in fact there could be such states.

In case of the hybrid model, this hypothesis is supported with different estimates of expected values of distribution parameters for different states of the hidden Markov chain. Were those estimates the same for all states, the hidden Markov chain would play no role in the model. Theoretically, the following situation seems possible: the neural network for one state could provide such a good fit of the model that the second state would not be needed any more. In the described situation the estimation results would deny the existence of a hidden Markov chain. Then it could be concluded that the model specification is incorrect and instead of the hybrid model, a neural network alone should be used. However, that is not a case in this research.

On the basis of the most likely sequences of states of the hidden Markov chain, it is shown that the state of lower seismic activity is a dominating one and periods of both higher and lower seismic activity in Poland are generally long.

Although the estimates are similar for both models, the hybrid model fits the data better than the HMM what demonstrates that hybrid HMM/ANN models can be an efficient alternative for HMMs. The advantage of the hybrid models over the HMMs consists in their larger flexibility and thus possibilities of getting a better fit. That flexibility makes it possible, among other analyses, to develop models of nonstationary time series. Obviously, the more complicated architecture, the larger flexibility. However, it seems that applications of the hybrid models with more complicated architectures might be limited with necessity of using extremely long time series. Moreover, estimation of a huge number of parameters might be difficult to perform in practice.

References

1. Baldi P., Brunak S., (2001). *Bioinformatics. The Machine Learning Approach*, Cambridge, MIT Press.
2. Barndorff-Nielsen O.E., Cox D.R., Klüppelberg C. (eds.), (2000). *Complex Stochastic Systems*, Boca Raton, Chapman and Hall.
3. Baum L.E., Petrie T., (1966). Statistical inference for probabilistic functions of finite Markov chains, *Annals of Mathematical Statistics*, 37, 1554-1563.
4. Elliott R.J., Lakhdar A., Moore, J.B., (1995). *Hidden Markov models: estimation and control*, New York, Springer.
5. Greene W.H., (2000). *Econometric Analysis*, Upper Saddle River, Prentice Hall.
6. Huang X.D., Ariky Y., Jack M.A., (1990). *Hidden Markov models for speech recognition*, Edinburgh, Edinburgh University Press.
7. Kurimo M., (1997). *Using self-organizing maps and learning vector quantization for mixture density hidden Markov models*, Helsinki, Helsinki University of Technology.
8. MacDonald I.L., Zucchini W., (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*, London, Chapman and Hall.
9. Rabiner L., (1989). A Tutorial on hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, 77(2), 257-285.
10. Riis S.K., Krogh A., (1997). Hidden neural networks: a framework for HMM/NN hybrids, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3233-3236.
11. Rynkiewicz J., (1999). Hybrid HMM/MLP models for time series prediction, *Proceedings of European Symposium on Artificial Neural Networks*, 455-462.
12. Rynkiewicz J., (2001). Estimation of Hybrid HMM/MLP models, *Proceedings of European Symposium on Artificial Neural Networks*, 383-390.
13. Trentin E., Gori M., (2001). A survey of hybrid ANN/HMM models for automatic speech recognition, *Neurocomputing*, 37(1-4), 91-126.
14. Zwoliński Z. *Wirtualna Geomorfologia*, http://main.amu.edu.pl/~sgp/gw/tzpl/gwtzpl.html