Δελτίο της Ελληνικής Γεωλογικής Εταιρίας τομ. XXXX, 2007
Πρακτικά 11^{ου} Διεθνούς Συνεδρίου, Αθήνα, Μάιος 2007

Bulletin of the Geological Society of Greece vol. XXXX, 2007
Proceedings of the 11th International Congress, Athens, May, 2007

IDENTIFICATION OF TEMPORAL PATTERNS IN THE SEISMICITY OF SUMATRA USING POISSON HIDDEN MARKOV MODELS

Orfanogiannaki K.[1], Karlis D.[2], and Papadopoulos G. A.[1]

[1] *National Observatory of Athens, Institute of Geodynamics, Lofos Nymfon, 11810, Athens, kath_orf@gein.noa.gr, g.papad@gein.noa.gr*

[2] *Athens University of Economics and Business, Department of Statistics, Patision 76Str., Athens, karlis@aueb.gr*

Abstract

On 26 December 2004 and 28 March 2005 occurred two of the largest earthquakes of the last 40 years between the Indo-Australian and the southeastern Eurasian plates, with moment magnitudes Mw=9.1 and Mw= 8.6 respectively. Complete data (mb ≥ 4.2) of the post-1993 time interval (Fig. 1) have been used to apply Poisson Hidden Markov Models (PHMM in identifying temporal patterns in the time series of the two main shocks. Each time series consists of earthquake counts, in given and constant time units, in the regions determined by the aftershock zones of the two main shocks. In PHMM each count is generated by one of m Poisson processes, that are called states. The series of states is unobserved and is, in fact a Markov chain. The model incorporates a varying seismicity rate; it assigns a different rate to each state, and detects the changes of the rate over time. In PHMM, unobserved factors related to the local properties of the region, affect the earthquake occurrence rate. Estimation and interpretation of the unobserved sequence of states that underlie the data contribute to a better understanding of the geophysical processes that take place in the region. We applied PHMM to the time series of earthquakes preceding the two main shocks, and we estimated the unobserved sequences of states that underlie the data. The results showed that the region of the 26 December 2004 earthquake was in state of low seismicity during about 400 days before the earthquake occurrence. On the contrary, in the region of the 28 March 2005 earthquake a transition from a state of low seismicity to a state of high seismicity was observed immediately after the occurrence of the big earthquake of 26 December 2004.

Key words: *Seismicity rate, Markov chains, Hidden states.*

Περίληψη

Στις 26 Δεκεμβρίου 2004 και στις 28 Μαρτίου 2005 σημειώθηκαν δύο από τις ισχυρότερες σεισμικές δονήσεις των τελευταίων 40 χρόνων ανάμεσα στη Ινδο-Αυστραλιανή και στην βορειοανατολική Ευρασιατική πλάκα, με μεγέθη ροπής Mw=9.1 και Mw= 8.6,αντίστοιχα. Πλήρη δεδομένα (mb ≥ 4.2) της μετά το 1993 χρονικής περιόδου χρησιμοποιήθηκαν για την εφαρμογή των Λανθάνοντων Μοντέλων Μάρκοβ με σκοπό την αναγνώριση των φάσεων σεισμικότητας στις χρονοσειρές των δύο ισχυρών σεισμών. Η κάθε χρονοσειρά αποτελείται από τις μηνιαίες συχνότητες των σεισμών που σημειώθηκαν στις μετασεισμικές ζώνες των δύο κύριων σεισμών. Ο καθορισμός των με-

τασεισμικών ζωνών των δύο ισχυρών σεισμών βασίστηκε στη χωρική κατανομή των μετασεισμών τους. Στα Λανθάνοντα μοντέλα Μάρκοβ η κάθε παρατήρηση παράγεται από μία από m κατανομές Poisson οι οποίες οναμάζονται καταστάσεις. Η χρονοσειρά των καταστάσεων είναι μή παρατηρούμενη και στην πραγματικότητα αποτελεί μία Μαρκοβιανή αλυσίδα. Το μοντέλο ενσωματώνει μεταβαλλόμενο ρυθμό σεισμικότητας, αντιστοιχεί διαφορετικό ρυθμό σεισμικότητας σε κάθε κατάσταση και αναγνωρίζει τις μεταβολλές του ρυθμού σεισμικότητας στο χρόνο. Στα Λανθάνοντα μοντέλα Μάρκοβ, μή παρατηρούμενοι παράγοντες που σχετίζονται με τις τοπικές ιδιότητες της περιοχής, θεωρούνται ότι επενεργούν στο ρυθμό σεισμικότητας. Η εκτίμηση και η ερμηνεία της μή παρατηρούμενης ακολουθίας των καταστάσεων που υπόκεινται των δεδομένων συμβάλλουν στην καλύτερη κατανόηση των Γεωφυσικών διαδικασιών που λαμβάνουν χώρα σε μία περιοχή. Στην εφαρμογή μας εφαρμόσαμε τα Λανθάνοντα Μοντέλα Μάρκοβ στις χρονοσειρές των δύο ισχυρών σεισμών και εκτιμήσαμε την ακολουθία των καταστάσεων που υπόκεινται των δεδομένων. Τα αποτελέσματα που εξήχθησαν έδειξαν ότι για περίπου 400 μέρες πρίν από τον ισχυρό σεισμό του Δεκεμβρίου η περιοχή βρισκόταν σε κατάσταση χαμηλής σεισμικότητας. Αντιθέτως, στην περιοχή του ισχυρού σεισμού του Μαρτίου παρατηρήθηκε μεταβολή στην κατάσταση σεισμικότητας, από κατάσταση χαμηλής σεισμικότητας σε κατάσταση υψηλής σεισμικότητας, αμέσως μετά τον ισχυρό σεισμό του Δεκεμβρίου.

Λέξεις κλειδιά: *Ρυθμός σεισμικότητας, Μαρκοβιανή αλυσίδα, μή παρατηρούμενες καταστάσεις.*

1. Introduction

Seismic events do not occur at regular time intervals, making the use of standard time series rather difficult. Time series with zeros can not be analysed with standard time series. An idea that is used in practice for the detection of temporal seismicity variations, is to count the number of events in a given time period, e.g. one month, and then to examine the resulting series. The Poisson distribution is the most adequate one to describe counts. One of the most widely known Poisson properties is that the mean of the counts equals the variance. In some cases, however, the mean is greater than the variance and the data are ovberdispersed. It is known that Poisson Mixture Models (PMM) can be applied to overdispersed heterogeneous data (McLachlan and Peel 2000, Titterington *et al.* 1985). However, data collected from the same area in successive time intervals tend to be dependent and, therefore, appropriate models for statistical modeling must accommodate this dependent structure. A class of models that allows for dependence between the data in addition to overdispersion is that of PHMM. The PHMM are extensions of the well known PMM and they decay to PMM in case of independent observations.

In PHMM each observation is generated by one of m Poisson distributions, called states. These states are unobserved (i.e. cannot be observed directly), hence the name PHMM. Each state has a different seismicity rate, while the series of states is in fact a Markov chain. Which state will generate the next observation depends on which state generated the current observation, through the transition probability matrix of the Markov chain. PHMM allow us to estimate the unobserved sequence of states that underlie the observation sequence. In this way we may reveal unknown properties of the mechanism that generated the data, and classify the observations with precision and objectivity. A recent PHMM application in identifying seismicity patterns can be found in (Orfanogiannaki 2006).

PHMM do not assume a constant rate for a long period of time. They incorporate a varying seismicity rate which is more realistic than a long-term constant rate. In fact, when a long-term constant rate is assumed short-term variations in seismicity are disregarded, although short-term variations in seismicity are important for the evaluation of the seismic activity in a region. PHMM assign a particular rate to each state. In this way, observations are classified according to the rate based on the rate that was determined for the previous observation. Additionally, changes to seismicity rate are detected.

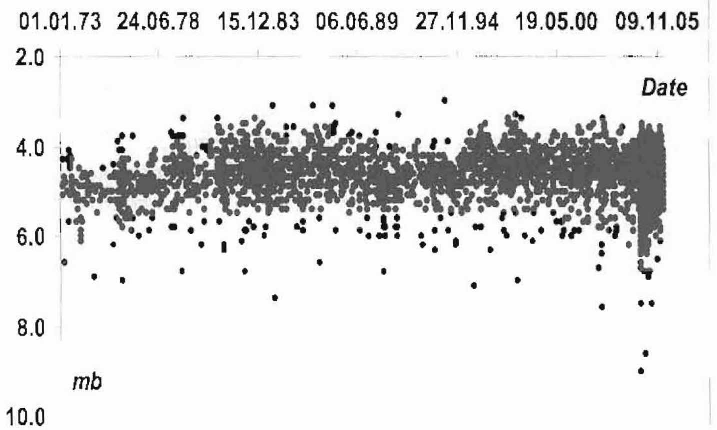

Figure 1 - Time-magnitude relationship for events occurring in the entire region examined from 01.01.1973 to 14.03.2006. The cut –off magnitude for completeness has been selected equal to *mb* = 4.2 for the post-1993 time interval

2. Data

The data sources are the USGS and ISC earthquake data files for the region E defined by the rectangle with coordinates -1.00°N–15.00°N and 91.00°E –100.00°E. At first, this region was divided into two sub-regions, *N* and *S*, based on the rapture zones of the two big earthquakes of 26.12.04 and 28.03.05, respectively (Lay *et al.* 2005). The solid line in the map (Fig. 2), shows the boundary between these two regions; the black stars correspond to the epicenters of the two main shocks. According to geophysical evidence, the rapture in the sub-region *N* was not uniform (Ammon *et al.* 2005). The rapture started in the southern part of the region and then propagated to the north. Based on the progress of the rapture, we divided sub-regiou *N* into two smaller regions N_1 and N_2 represented by the two dashed lines (Fig. 2). Data completeness analysis based on the magnitude-frequency relationship showed that the data in all regions are complete for $Mb \geq 4.2$ for the time interval from 1994 onwards. All data sets are actually discrete valued time series, since they count the number of events in twenty-three day time periods. This time unit was selected so as to have an integer number of periods covering the time between the two big earthquakes of 26.12.04 and 28.03.05.

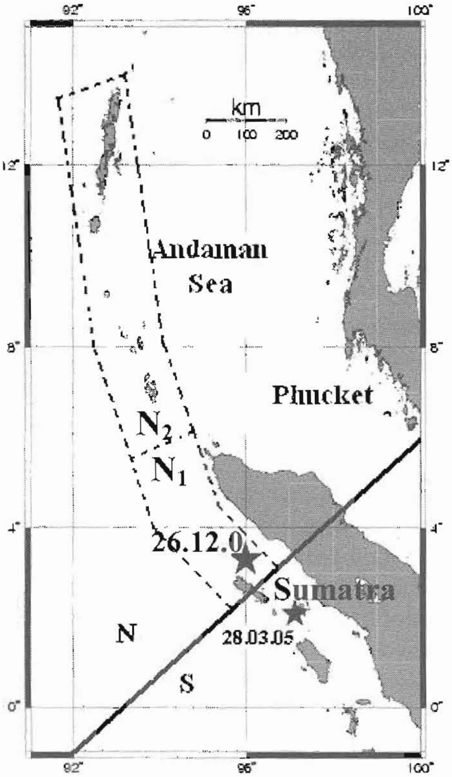

Figure 2 - Map of the area

3. Poisson hidden markov models: definition and notation

- PHMM are discrete time stochastic processes that consist of an unobserved finite state Markov chain $\{C_t: t \in N\}$ having m states and an observed sequence of a non-negative integer valued stochastic process $\{S_t: t \in N\}$ such that for all positive integers T, conditionally on $C^{(T)}=\{C_t{:}t=1,\ldots,T\}$ the random variables $S_1,\ldots,S_T$ are independent.
- The marginal distribution of S_t is:

$$p(s_t) = \sum_{j=1}^{m} a_j f(s_t \mid \lambda_j)$$

where $a_i > 0,\ \ i = 1,\ldots,m,\ \ \sum_{i=1}^{m} a_i = 1$ and $f(s \mid \lambda) = \frac{e^{-\lambda}\lambda^s}{s!}, s = 0,1,\ldots,\ \ \lambda \geq 0$

- The conditional distribution of S_t given $C^{(T)}$ is:

$$\pi_{s_t i} = P(S_t = s_t \mid C_t = i) = \frac{e^{-\lambda_i}\lambda_i^{s_t}}{s_t!},\ \ s_t = 0,1,\ldots,\ \ \lambda \geq 0$$

- The transition probabilities of the Markov chain are:

$$\gamma_{ij} = P(C_t = j \mid C_{t-1} = i)$$

i.e. γ_{ij} is the probability to move from state i, at time t-1, to state j at time t, for any states i,j=1,...,m and for any time t=1,...,T

4. Estimation of the unknown parameters

- Estimation of the parameters of interest is obtained via the EM-algorithm (Dempster *et al.* 1977). The EM-algorithm, though, may be significantly simplified using the "forward" $\alpha_t(i)$ and "backward" $b_t(i)$ probabilities introduced by Baum *et al.* (1970)(i=1,...,m, t=1,...,T).
- The "forward" probability $\alpha_t(i)$ is the joint probability of the past and present observations and the current state of the Markov chain:

$$\alpha_t(\mathrm{i}) = \mathrm{P}(\mathrm{s}_1,\ldots,\mathrm{s}_t,\mathrm{C}_t = \mathrm{i})$$

- The "backward" probability $b_t(i)$ is the conditional probability of the future observations given the current state of the Markov chain:

$$b_t(\mathrm{i}) = \mathrm{P}(\mathrm{s}_{t+1},\ldots,\mathrm{s}_T \mid \mathrm{C}_t = \mathrm{i})$$

The computation of the "forward" and "backward" probabilities is based on recursive algorithms (Leroux *et al.* 1992).

- The indicator random variables $u_j(t)$ and $\upsilon_{jk}(t)$, where $u_j(t)$=1, if C_t=j and 0 otherwise and $\upsilon_{jk}(t)$=1 if C_{t-1}=j and C_t=k are treated as missing data in the EM-algorithm. The EM-algorithm is an iterative algorithm that consists of two steps. If we denote as starting values the values $(\lambda_1,\dots,\lambda_m,\gamma_{11},\dots,\gamma_{mm})$, the two steps of the EM-algorithm are:

- E-step: Calculate $u_j(t)$ and $\upsilon_{jk}(t)$ using equations 1 and 2, respectively.

Equation 1 - Calculate $u_j(t)$

$$u_j(t) = \frac{a_t(j)b_t(j)}{\sum_{i=1}^{m} a_T(i)}$$

Equation 2 - Calculate $\upsilon_{jk}(t)$

$$\upsilon_{jk}(i) = \frac{a_{t-1}(i)b_t(j)\gamma_{ij}\pi_{s_t j}}{\sum_{i=1}^{m} a_T(i)}$$

- M-step: Update the estimates γ_{jk}, $i,j=1,\dots,m$, λ_j, $i=1,\dots,m$ using equations 3 and 4, respectively.

Equation 3 – Update γ_{jk}

$$\gamma_{jk}^{(new)} = \frac{\sum_{t=2}^{T} \upsilon_{jk}(t)}{\sum_{t=2}^{T}\sum_{i=1}^{m} \upsilon_{ji}(t)}$$

Equation 4 – Update λ_j

$$\lambda_j^{(new)} = \frac{\sum_{t=1}^{T} u_j(t)s_t}{\sum_{t=1}^{T} u_j(t)}$$

If the difference between the starting values and the new estimated values is less than 10^{-10}, stop iterating, otherwise set as starting values the new estimated values and go to the E-step.

- The estimation of the unobserved state C_t, at time t, that underlies the corresponding observed state S_t is based on the probability:

$$P(C_t = i \mid s_1,\dots,s_T) = \frac{a_t(i)b_t(i)}{\sum_{i=1}^{m} a_T(i)}$$

The state that maximizes the above probability consstitutes an estimate of C_t.

algorithm is an iterative algorithm that consists of two steps. If we denote as starting values the values $(\lambda_1,\ldots,\lambda_m,\gamma_{11},\ldots,\gamma_{mm})$, the two steps of the EM-algorithm are:

- E-step: Calculate $u_j(t)$ and $\upsilon_{jk}(t)$ using equations 1 and 2, respectively.

Equation 1 - Calculate $u_j(t)$

$$u_j(t)=\frac{a_t(j)b_t(j)}{\sum_{i=1}^{m} a_T(i)}$$

Equation 2 - Calculate $\upsilon_{jk}(t)$

$$\upsilon_{jk}(i)=\frac{a_{t-1}(i)b_t(j)\gamma_{ij}\pi_{s_t j}}{\sum_{i=1}^{m} a_T(i)}$$

- M-step: Update the estimates γ_{jk}, $i,j=1,\ldots,m$, λ_j, $i=1,\ldots,m$ using equations 3 and 4, respectively.

Equation 3 – Update γ_{jk}

$$\gamma_{jk}^{(new)}=\frac{\sum_{t=2}^{T}\upsilon_{jk}(t)}{\sum_{t=2}^{T}\sum_{i=1}^{m}\upsilon_{ji}(t)}$$

Equation 4 – Update λ_j

$$\lambda_j^{(new)}=\frac{\sum_{t=1}^{T}u_j(t)s_t}{\sum_{t=1}^{T}u_j(t)}$$

If the difference between the starting values and the new estimated values is less than 10^{-10}, stop iterating, otherwise set as starting values the new estimated values and go to the E-step.

- The estimation of the unobserved state C_t, at time t, that underlies the corresponding observed state S_t is based on the probability:

$$P(C_t=i\mid s_1,\ldots,s_T)=\frac{a_t(i)b_t(i)}{\sum_{i=1}^{m} a_T(i)}$$

The state that maximizes the above probability consstitutes an estimate of C_t.

5. Analysis

We applied PHMM to the entire region *E*, as well as to the sub-regions *S, N*, N_1 and N_2. The model selection for each region was based on the AIC information criterion (Akaike 1974). Once the model has been selected (i.e. the number of states was determined), the model parameters estimates for each region were obtained via the EM-algorithm (Section 4). The Poisson rates and the transition probability matrix are illustrated in Table 1. Additionally, the unobserved sequence of states that underlie the data was estimated for each region.

Table 1- Model Parameters estimates

Segment	Number of components	Component Number i	Aic	Log-likelihood	Parameters estimates	
					Poisson Rates λi	Transition Probability matrix
E	4	1	707.7	-337.877	4.85	$\left(\begin{matrix} 0 & 0.963 & 0 & 0.03 \\ 0.465 & 0.526 & 0 & 0.00 \\ 0.104 & 0 & 0.658 & 0.23 \\ 0 & 0 & 1 & 0 \end{matrix}\right.$
		2			8.86	
		3			9.61	
		4			28.20	
S	3	1	344.0	-163.016	1.78	$\begin{pmatrix} 0.966 & 0.016 & 0.018 \\ 0.133 & 0.689 & 0.178 \\ 0 & 0.573 & 0.427 \end{pmatrix}$
		2			4.13	
		3			13.15	
N	3	1	425.9	-203.985	1.46	$\begin{pmatrix} 0.490 & 0.424 & 0.086 \\ 0.406 & 0.495 & 0.099 \\ 0.213 & 0.787 & 0 \end{pmatrix}$
		2			5.21	
		3			19.42	
N1	2	1	298.9	-145.493	1.38	$\begin{pmatrix} 0.911 & 0.089 \\ 0.460 & 0.540 \end{pmatrix}$
		2			5.17	
N2	3	1	347.5	-164.782	1.29	$\begin{pmatrix} 0.742 & 0.206 & 0.052 \\ 0.803 & 0 & 0.197 \\ 0.334 & 0.666 & 0 \end{pmatrix}$
		2			4.15	
		3			18.15	

The application of PHMM to the complete data set for the entire region examined showed (Fig. 3) that the state of seismicity ranges only from 1 to 2 in the interval 1994 – 2002; that is, the seismicity is relatively low. From 2002 onwards a transition to higher states of seismicity is observed; that is, to states 3 and 4, with rates 9.61 and 28.20 (events/23days), respectively (Table 1). To emphasize on the period of the increased seismicity, we narrow the time window examined. The seismicity states were further investigated for the time interval 2000 – 25.12.2004 (inclusive) for sub-regions *N*, N_1 and N_2, as well as for the time interval 1.1.2000 – 27.3.2005 (inclusive) for sub-region *S*.

Sub-region *S* was characterized by state 2 before the big earthquake of 26.12.04 which occurred in sub-regions N_1 (Fig. 4b). However, a transition from state 2 to state 3 of high seismicity was observed immediately after the occurrence of the big earthquake of 26.12.04. This may imply a triggering effect due to stress transfer from N_1 to *S*.

In sub-region N_1 no state of high seismicity is observed (Fig. 4d) before the big earthquake of 26.12.04. On the contrary, during about 400 days before the earthquake occurrence the state of seismicity is 1, that is, low seismicity prevails. The low seismicity observed in sub-region N_1 is due to the fact that only one strong earthquake occurred in the time interval examined. This event though was deep and was not followed by aftershocks which would increase the seismicity in the region. In sub-region *N* as well as in sub-region N_2, the state 3 of high seismicity appears at some certain points of time (Figs 4a, c), and these are attributed to aftershock activity associated with strong earthquake

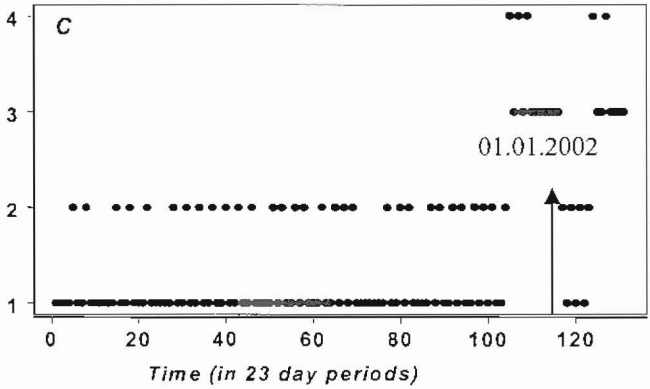

Figure 3 - Estimated states, *C*, that underlie the data against time (in 23-day periods) for the entire region (*E*) examined. The zero point of time is 01.01.1994

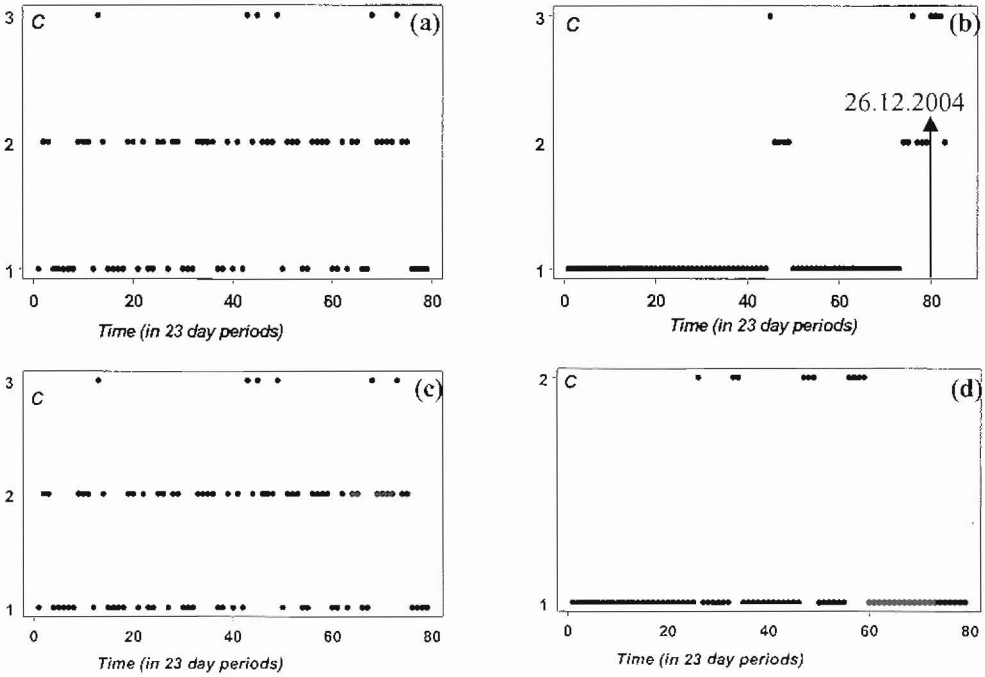

Figure 4 - Estimated states, *C*, that underlie the data against time (in 23-day periods): (a) sub-region *N*, (b) sub-region *S*, (c) sub-region N_2 and (d) sub-region N_1. The zero point of time is 01.01.2000 for all the sub-regions examined

6. Conclusions

PHMM provide a diagnostic tool for identifying changes in seismicity states. The model incorporates a varying seismicity rate, detects the changes on the rate over time, and assigns a particular rate for each state. Estimation of the sequence of unobserved states that underlie the data is attained with relative easiness. In the region of the 26.12.04 earthquake during about 400 days before the earthquake occurrence the state of seismicity is 1; that is, low seismicity prevails. On the contrary, in the region of the 28.03.05 earthquake, before it occurred, a transition from state 2 to state 3 of high seismicity was observed immediately after the occurrence of the big earthquake of 26.12.04. Our analysis was based on the assumption that the time unit in which we count the number of events is fixed. It would be interesting to examine how the selection of alternative time units can change the estimated patterns.

7. References

Akaike, H., 1974. A new look at the statistical model identification, *IEEE Transactions on Automaric Control*, AC-19, 716-723.

Ammon, C.J., Ji, C., Thio, H.-K., Robinson, D., Ni, S., Hjorleifsdottir, V., Kanamori, H., Lay, T., Helmberger, D., Ichinose, G., Polet, J., and Wald, D., 2005. Rupture Process of the 2004 Sumatra-Andaman earthquake, *Science*, 308, 1133-1139.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N., 1970. A maximization technique in the Statistical analysis of probabilistic functions of Markov chains, *The Annals of Mathematical Statistics*, 41, 164-171.

Dempster, A.P., Laird, N.M., and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the the EM algorithm (with discussion), *Journal of the Royal Statistical Society B*, 39, 1-38.

Lay, T., Kanamori, H., Ammon, C.J., Nettles, M., Ward, S.N., Aster, R.C., Beck, S.L., Bilek, S.L., Brudzinski, M.R., Butler, R., DeShon, H.R., Ekström, G., Satake, K., and Sipkin, S., 2005. The great Sumatra-Andaman earthquake of 26 December 2004, *Science*, 308, 1127-1133.

Leroux, B.G., and Puterman, M.L., 1992. Maximum –penalized likelihood estimation for independent and Markov-dependent mixture models, *Biometrics*, 48, 545-558.

Orfanogiannaki, K., 2006. Identification of seismicity patterns using Hidden Markov Models, *Master Thesis*, Athens Univercity of Economics and Business, Athens, 105pp.