

시계열 분석에 의한 어획량 예측

— 한국 근해산 갈치를 예로 하여 —

유신재 · 장창익*

한국해양연구소 · *군산대학교 해양개발학과

Forecasting of Hairtail (*Trichiurus lepturus*) Landings in Korean Waters by Times Series Analysis

Sinjaee YOO and Chang-Ik ZHANG*

Korea Ocean Research & Development Institute, Ansan P. O. Box 29, 425-600 Korea

**Department of Marine Development, Kunsan National University,*

Kunsan, 537-360 Korea

Short-term forecasting of fish catch is of practical importance in fisheries management. Ecosystem models and multi-species models as well as traditional single-species models fall short of predicting power needed for practical management of fisheries resources due to the lack of sufficient data or information for the required parameters. Univariate time series analysis, on the other hand, extracts the information on the stochastic variability from the time series itself and makes estimates of the future stochastic variability. Therefore, it can be used for short-term forecasting with minimum data requirements.

ARIMA time series modeling has been applied to the monthly Korean catches of hairtail (*Trichiurus lepturus*) for 1971~1988. Forecasts of hairtail catch were made and compared with the actual catch data from 1989~1990 which were not included in the parameter estimation. The results showed a good agreement ($r=0.938$) between the forecasts and the actual catches with a mean relative error of 59.5%

서 론

수산학에서 자원관리와 어획예측은 자원평가와 더불어 가장 중요한 사항들이다. 특히 어획예측은 관리를 행하는데 있어서 중심되는 역할을 한다 (Makridakis *et al.*, 1983). 정확한 예측은 자원생산, 예산 및 정책결정 등 여러가지의 불확실성으로 인해 어려운 과제이다 (Hilborn, 1987). 특히 수산 자원관리에 있어 중요한 정보인 어획량의 예측에는 아직 만족할 만한 결과를 낳지 못하고 있다. 어획량은 해양학적, 생물학적 요인뿐 아니라 사회, 경제적 제 요인에 의한 복합적 영향을 받기 때문이다.

수산학에서 사용되어 오고 있는 전통적인 개체

군역학 모델에는 이러한 복합적 요인중 종내경쟁과 어획사망을 등만이 명시적으로 포함되며, 나머지 생태학적 요인들은 자연사망율이나 개체군 성장을 등의 주요 매개변수 안에 암시적으로 포함되어 있다. 복합적 요인의 영향은 명확히 정의되어 있지 않고 이들 요인 각각의 기여도는 파악하기가 어렵다. 자연 생태계에서 이들 요인의 개체군 변동에 대한 기여도는 시·공간적으로 뿐만 아니라, 종별로도 크게 변할 것으로 생각된다. 따라서 단일 개체군역학 모델은 제한된 범위의 요인들만을 다루게 되는 한계성 때문에 적용범위와 유용성에 있어서 제한이 된다.

종간 경쟁이나 포식-피식 관계 등 생물 간의

상호작용을 명시적으로 다루고 있는 다중 모델과, 더 나아가 물리적 요인까지 포함시키는 생태계 모델들은 개체군의 변동 예측과 같은 실제적 응용 분야에서는 아직까지 시도의 단계에 국한되고 있다. 다중 모델이나 생태계 모델이 예측에 있어 크게 성공적이지 못한 것은 대체적으로 두 가지 이유에서라고 할 수 있다. 첫째, 모델이 복잡해짐에 따라 변수와 더불어 매개변수의 수가 늘어나게 되는데 대개의 경우 이러한 매개변수들의 측정이나 추정 이 실제적으로 어렵고, 그 결과 모델 자체의 불확실성이 커지게 된다. 두번째 이유는 이들 모델의 변수나 매개변수들이 결정론적(deterministic)으로 정의가 되기 때문이다. 자연계에서는 이들 변수 또한 확률론적으로 변동하게 되는데 다중 모델인 경우 매개변수의 평균값 추정도 어려운 경우가 많으므로 매개변수의 분산은 더욱 추정하기가 어렵다. 이러한 두 가지 이유에서 다중 모델이나 생태계 모델은 생태계 변화에 대한 정량적 예측에 있어 아직은 실용성이 적으며 앞으로 계속적인 연구에 의한 개선의 여지가 크다.

시계열 분석(time series analysis)은 시계열 내의 값들이 가지는 상관관계를 분석하여 시계열의 변동형태를 파악하고 이를 토대로 예보(forecasting)를 하는 방법이다. 이때 시계열 외의 다른 자료는 필요치 않으므로 제 요인에 대한 충분한 정보가 결여된 조건에서도 분석이 가능하며, 특히 단기간의 예보는 증권시장 등 경제분야에서 효과적인 방법으로 응용되어 왔다. 수산자원 관리에서도 응용이 되고 있으나(Saila *et al.*, 1977; Stergiou, 1989, 1990) 몇몇 연구에 그치고 있다. 예보에 있어 특히 많이 쓰이는 모델은 단변수(univariate) 방법인 ARIMA(Auto Regressive Integrated Moving Average) 모델이다(Box and Jenkins, 1976).

본 연구는 1970년에서 1988년까지의 한국연근해에서 어획된 갈치의 월별 어획량 자료를 ARIMA model로 분석하여 2년 간의 예보를 하고 이 예보 값들과 1989년, 1990년의 실제 어획량을 비교하였다.

자료 및 방법

1. 자 료

본 연구에 사용된 자료는 한국연근해에서 어획된 갈치의 1970~1990까지의 21년 간에 걸친 월별 어획량 자료이다. 갈치의 월별 어획량은 수산통계

연보(1970~1983), 어업생산량통계(1984, 1990), 농림수산통계연보(1985~1990)로부터 구하였다. 어획량 자료는 비정상성 분산(nonstationary variance)을 줄이기 위하여 대수변환(logarithmic transformation)을 하여 분석하였다. 자료 분석에는 컴퓨터 통계패키지인 SYSTAT 4.0(Wilkinson, 1988)을 사용하였다.

2. ARIMA 모델

한 시점의 변량은 그 이전의 변량들의 영향을 받거나(autoregressive), 또는 과거의 오차(error term)의 영향을 받고 있다고 볼 수 있다(moving average). 자기회귀(Autoregressive) 모델은 다음과 같다.

$$x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \dots + \Phi_p x_{t-p} + e_t \quad \dots\dots (1)$$

즉, 변량 x_t 는 p 개의 과거값의 선형조합(linear combination)과 무작위적 오차항 e_t 의 합으로 표현할 수 있다. e_t 는 평균이 0이며, 분산은 σ_e^2 이다. 이 모델은 p 차 AR 모델이라고 부른다. 후진 연산자(back-shift operator) B 를 써서 식(1)을 간략히 표현하면,

$$(1 - \sum_{k=1}^p \Phi_k B^k) x_t = e_t$$

$$\Phi(B) x_t = e_t$$

$$\text{여기서, } \Phi(B) = 1 - \Phi_1 B - \dots - \Phi_p B^p$$

한편, 이동평균(moving average) 모델은 x_t 가 무한대적인 e 의 평균합이 아니라 한정된 수 q 개의 이전 e 들에 종속된다는 가정하에서 다음과 같이 나타낸다.

$$x_t = e_t + \beta_1 e_{t-1} + \dots + \beta_q e_{t-q} = \theta(B) e_t$$

$$\text{여기서, } \theta(B) = 1 + \beta_1 B + \dots + \beta_q B^q \quad \dots\dots\dots (2)$$

따라서, 두 과정을 합친 모델은 자기회귀이동평균(ARMA) 모델이라 부르는데,

$$x_t = \Phi_1 x_{t-1} + \dots + \Phi_p x_{t-p} + e_t + \beta_1 e_{t-1} + \dots + \beta_q e_{t-q}$$

이를 정돈하면,

$$\Phi(B) x_t = \theta(B) e_t$$

로 표현된다.

시계열 분석은 정제 시계열(stationary time series)을 가정하지만, 실제의 많은 시계열이 이에 어긋나므로 대개 시계열 점들의 차를 구하여 새 시계열을 만들 경우 정체성이 만족되게 된다. 즉,

$$W_t = x_t - x_{t-d} = \nabla^d x_t$$

이 새 시계열 W_t 에 ARMA 모델을 적용할 때, 이를

ARIMA(autoregressive integrated moving average) 모델이라 한다. 계절적 패턴이 있을 경우도 포함하면 ARIMA 모델의 일반식은

$$\Phi_p(B)\Phi_q(B^{12})W_t = \theta_q(B)\theta_p(B^{12})e_t \quad \dots\dots\dots (3)$$

W_t 는 계절적, 순차적 차 계산(differencing)을 한 값이다.

$$W_t = \nabla^d \nabla_{12}^D x_t$$

식 (3)은 간략히 ARIMA(p, d, q)(P, D, Q)₁₂로 표현하기도 한다.

ARIMA 모델링의 과정은 모델 동정(model identification), 매개변수 추정(parameter estimation) 및 진단(diagnosis)으로 나눌 수 있다. 모델 동정이란 AR, MA, 계절성 등의 차수, 즉, p, q, d, P, Q, D 등을 결정하는 과정이다. AR, MA는 자기상관함수(ACF: Auto Correlation Function)와 편자기상관함수(PACF: Partial Auto Correlation Function)에 있어 각각 특징적 양상을 보이므로 시계열의 ACF, PACF를 분석하여 AR, MA의 차수를 결정하게 된다(Box and Jenkins, 1976). 시계열이 정체성이 없을 경우 계절적, 비계절적 차감(differencing)을 하는데 대개 1, 2차로 만족스런 결과를 얻을 수 있다. 매개변수의 추정은 결정된 모델을 토대로 시계열 값을 적합시켜 이루어진다. 매개변수가 추정되면 모델의 적합성을 판정하게 되는데, 그 방법으로는 첫째, 잔차의 ACF, PACF를 점검하고, 둘째, 잔차의 누적주기도(integrated periodogram)를 통하여 편차를 점검하고, 셋째, 매개변수 추정에 쓰이지 않은 실제값과 모델에서 예보된 값을 비교하는 것 등이 있다(Shumway, 1988).

결 과

Fig. 1은 대수변환한 값치의 연도별 월별 어획량이다. 어획량의 변동은 연도에 따라 진폭은 다르나 계절에 따르는 주기를 뚜렷이 나타내고 있다. 이에 대해 계산된 ACF와(Fig. 2) PACF를(Fig. 3) 보면 12개월마다 유의한 값이(95% 신뢰구간보다 큰 값) 되풀이되며 천천히 줄어드는 계절적 비정체성을 나타내고 있으므로 계절적 차감(seasonal differencing)이 필요하다고 판단된다. 따라서 먼저 계절적 차감(D=1)을 하여 ACF, PACF를 계산해 보면 Fig. 4와 Fig. 5와 같다. 일반적으로 모델 동정에 있어 유의한 ACF는 MA항이 있음을, PACF는 AR항이 있음을 의미한다(Shumway, 1988). 따라서

Fig. 4의 lag=1, 2, 12에서 보이는 유의한 값을 없애기 위해 계절적 MA를 포함시키고 Fig. 5의 lag=1에 있는 유의한 값을 없애기 위해 비계절적 AR를

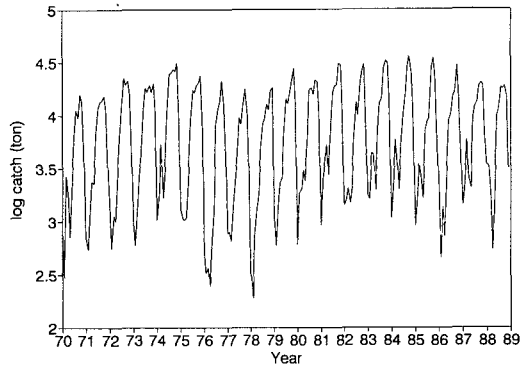


Fig. 1. Log-transformed (Base=10) time series of the monthly catch of hairtail.

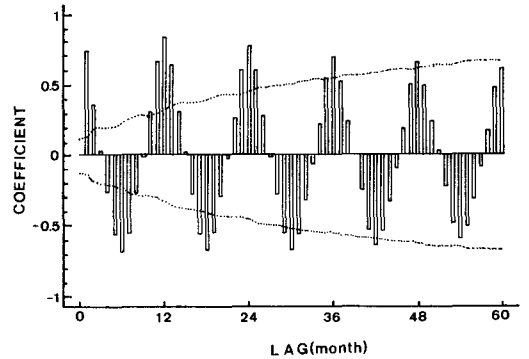


Fig. 2. Estimated ACF of the log-transformed hairtail catch shown with 95% confidence interval.

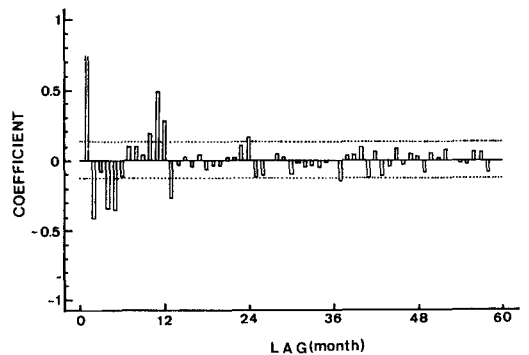


Fig. 3. Estimated PACF of log-transformed hairtail catch shown with 95% confidence interval.

넣었다[ARIMA(1,0,0) (0,1,1)]. 이 모델의 잔차(residual)를 분석한 것이 Fig. 6과 Fig. 7이다. 여기서 ACF, PACF 다 95%의 신뢰구간을 벗어나는 유의한 값이 없으며 이는 즉 모델이 적합함을 의미한다.

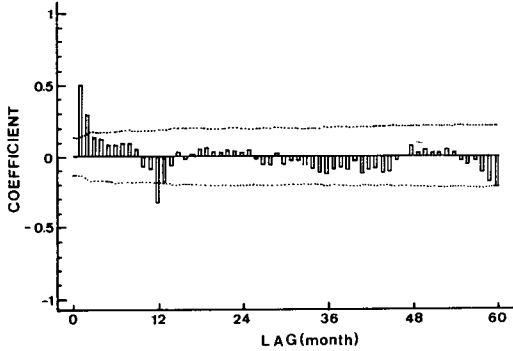


Fig. 4. Estimated ACF after seasonal differencing is made to hairtail catch.

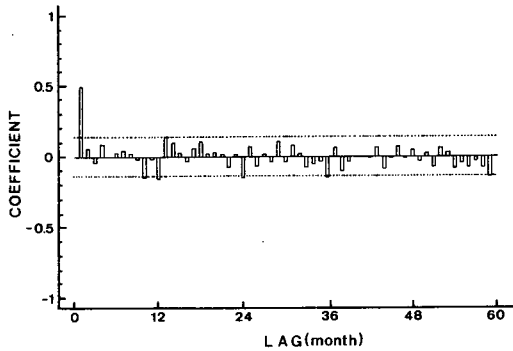


Fig. 5. Estimated PACF after seasonal differencing is made to hairtail catch.

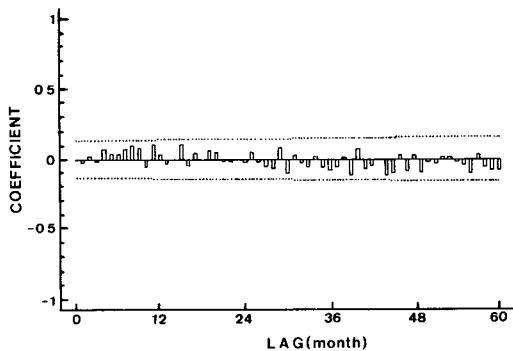


Fig. 6. Estimated residual ACF of ARIMA(1,0,0) (0,1,1)₁₂ on hairtail catch.

Box와 Jenkins(1976)는 계절적 경향성을 검증하는 방법으로 ACF보다 더 민감한 Bartlett(1955)의 누적 주기도(cumulative periodogram)를 쓸 것을 권하고 있다. 잔차에 대한 누적 주기도 역시 이 모델이 적합함을 보여준다(Fig. 8). 누적주기도에서는 표준화된 값의 누적분포를 보여주는데 모든 값들이 95% Kolmogorov-Smirnov 경계안에 들어가고 있다. 이는 잔차들이 무작위적 분포를 하고 있음을, 즉 모델이 자료가 보여주는 현상을 성공적으로 재현함을 의미한다. ARIMA(1,0,0) (0,1,1)₁₂을 다시 표현하면,

$$(1 - n_1B)(1 - B^{12})x_t = (1 - U_1B^{12})e_t$$

이는 $x_t = n_1x_{t-1} + x_{t-12} - n_1x_{t-13} + e_t - U_1e_{t-12}$ 이고, 이 식을 적합하여 Table 1과 같은 결과를 얻었다. 즉 추정된 모델식은 다음과 같다.

$$x_t = 0.546x_{t-1} + x_{t-12} - 0.546x_{t-13} + e_t - 0.562e_{t-12}$$

위의 모델식을 토대로 향후 2년 간의 갈치의 월

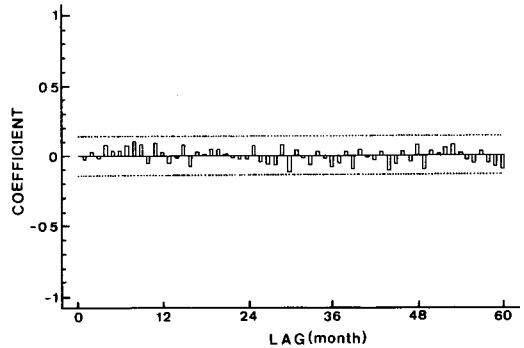


Fig. 7. Estimated residual PACF of ARIMA(1,0,0) (0,1,1)₁₂ on hairtail catch.

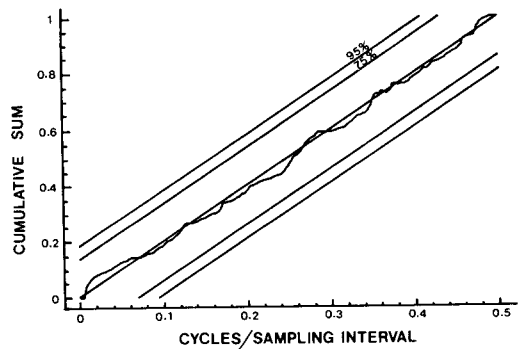


Fig. 8. Cumulative periodogram for residuals of ARIMA(1,0,0) (0,1,1)₁₂ on hairtail catch.

별 어획량을 예보한 것이 Fig. 9이다. 그림에서 보는 바와 같이 예측치는 실측치에 잘 적합되었다. 상관계수는 0.938이었으며 평균 상대오차는 59.5%로 추정되었다. 첫해의 예측치는 상대적으로 실측치에 가까웠으나 둘째해의 예측치는 첫해에 비해 덜 일치하는 경향을 보였다.

Table 1. Parameter estimation of ARIMA(1,0,0) (0,1,1)₁₂ of hairtail catch.

Parameter	Estimate	Std. error	t-value	P-value
n ₁	0.54558	0.05798	9.40944	0.00000
U ₁	0.56218	0.06172	9.10827	0.00000

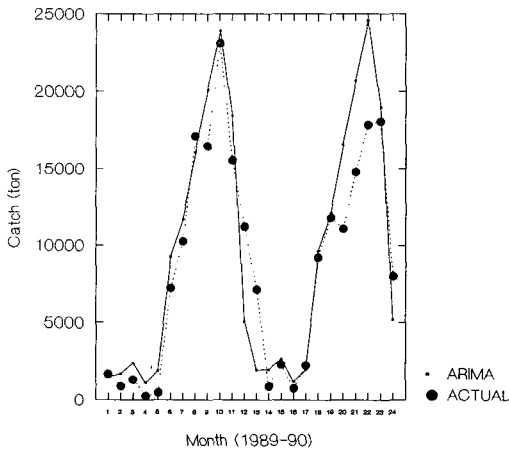


Fig. 9. Comparison of forecasted catch and actual catch.

고찰

ARIMA 모델링의 목표는 생물학적, 물리적, 사회경제적 변동요인의 불확실성에도 불구하고 단기적 예측을 하는 것이다. 일반적으로 단기 예측이란 9개월 내의 예측을 말한다(Chatfield, 1984). 본 연구에서 사용된 모델에 의해 예측된 예측치와 실제치가 전반적으로 잘 일치함을 알 수 있다. 그러나 상대오차를 구해보면 전체적으로 약 59.5%의 큰 값을 보인다(오차범위 = 2.6~339%). 이 값은 Stergiou(1990)가 그리이스의 Mullidae 어획량을 ARIMA 모델로 분석예측해서 얻은 12%보다 훨씬 크다. 이 원인은 갈치의 어획량이 가지는 값의 범

위가 Mullidae보다 훨씬 크다는데 있다. 갈치의 경우 예측기간 동안의 실제값의 최대값은 최소값의 약 100배였으나, Mullidae는 약 3배의 작은 범위를 보였다. 특히 갈치의 경우, 초기의 값들이 매우 작아 비교적 작은 절대오차에도 불구하고 큰 상대오차를 보였다. 이러한 초기값을 제외하면 상대오차의 평균값은 약 31.7%로 떨어지게 된다.

한편 예보기간의 후반, 즉 2년째의 예측은 과대평가하고 있다. 과거 20년 간 갈치자원은 장주기의 진동을 하는 것으로 보이는데(해양연구소, 1991), 주기가 일정하지 않으므로 시계열 모델로는 이를 예측하기가 어렵다.

본 연구를 통해 분석된 결과는 ARIMA 모델이 가지는 장점을 잘 보여준다. 즉, 단지 과거의 어획량 자료의 변동성만을 분석하여 기타의 생물학적, 해양학적, 경제적 정보없이도 비교적 정확한 단기 예측을 할 수 있다는 점이다. 이러한 장점에도 불구하고 시계열 분석법은 수산학 분야에서 현재 널리 사용되지 않고 있다. ARIMA 모델은 단변수 분석법이나 다변수 방법으로 전이함수모델(transfer function model)도 쓰이고 있다(Keller, 1987). 최근에는 벡터자기회귀(vector auto regression) 모델도 적용되고 있다(Stergiou, 1991).

요약

어획량의 단기 예측은 자원관리에 있어 중요한 항목이지만 전통적인 개체군 모델은 수산자원 관리에 있어 실제로 요구되는 예측력이 크게 부족하다. 다종 또는 생태계 모델도 요구되는 매개변수의 수가 많아 실제적 적용이 어렵다. 반면에 단변수 시계열 분석법은 시계열 자체에서 변동성에 관한 특성을 추정하여 이를 토대로 장래 변동성을 예측함으로써 최소한의 자료를 가지고 비교적 정확한 단기에측이 가능하므로 유용성이 높다.

본 연구에서는 ARIMA 시계열 모델을 1971~1988년 간의 한국근해의 월별 갈치어획량 자료에 적용하였다. 여기서 나온 예측치와 분석에 포함되지 않았던 1989~1990년 간의 어획량과 비교하였다. 분석 결과 예측치와 실제어획량이 잘 일치하였으며($r=0.938$) 평균상대오차는 59.5%였다.

참고문헌

- Box, G. E. P. and G. M. Jenkins. 1976. Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco.
- Bartlett, M. S. 1955. Stochastic Processes. Cambridge University Press.
- Chatfield, C. 1984. The analysis of time series: an introduction. Chapman and Hall, New York.
- Hilborn, R. 1987. Living with uncertainty in resource management. North Am. J. Fish. Magt. 7 : 1~5.
- Keller, A. 1987. Modeling and forecasting primary production rates using Box-Jenkins transfer function models. Can. J. Fish. Aquat. Sci. 44 : 1045~1052.
- Makridakis, S., S. Wheelwright and V. McGee. 1983. Forecasting: methods and applications. John Wiley & Sons, New York. 926pp.
- Saila, S. B., M. Wigbout and R. T. Lermitt. 1979. Comparison of some time series models for the analysis of fisheries data. J. cons. int. Explor. Mer. 39 : 44~52.
- Shumway, R. H. 1988. Applied statistical time series analysis. Prentice-Hall, 379pp.
- Stergiou, K. I. 1989. Modeling and forecasting the fishery for pilchard(*Sardina pilchardus*) in Greek waters using ARIMA time-series models. J. Cons. int. Explor. Mer. 46 : 16~23.
- Stergiou, K. I. 1990. Prediction of the Mullidae fishery in the eastern Mediterranean 24 months in advance. Fisheries Research 9 : 67~74.
- Stergiou, K. I. 1991. Describing and forecasting the sardine-anchovy complex in the eastern Mediterranean using vector autoregressions. Fisheries Research 11 : 127~141.
- Wilkinson, L. 1988. SYSTAT: The System for Statistics. SYSTAT, Inc. Evanston, IL.
- 해양연구소. 1991. 주요 수산자원의 합리적 관리를 위한 이론 모델과 소프트웨어의 개발. 과학기술처 BSPG 00121-372-3, 211pp.

1993년 2월 10일 접수

1993년 7월 3일 수리