http://dx.doi.org/10.17208/jkpa.2015.02.50.2.227 대한국토·도시계획학회지「국토계획」제50권 제2호 2015. 02

데이터 마이닝을 이용한 한국 가구 근로소득 분석에 관한 연구

A Study on Korean Household Income Using Data Mining

전해정*
Chun, Hae Jung

Abstract

The purpose of this study is to investigate the difference in information of the groups with high and low Korean household income before and after the global financial crisis, by the empirical analysis using empirical data mining with the data from the Korea Labor Panel. The predictive accuracy of the model determined by decision tree, logit regression and neural network is shown to be significantly high, with the highest in the predictive value of the logit regression.

According to the empirical results, savings is found to affect most on the difference between the high and low group in their Korean household income, followed by the type of home owning and renting. According to the difference analysis before and after the global financial crisis, the influence of whether they own real estates except houses on the differences between the earned income groups, after the global financial crisis due to the price fall.

키 워 드 ▪ 근로소득, 데이터 마이닝, 로지스틱 회귀분석, 의사결정나무, 신경망 분석
Keywords ▪ Income, Data Mining, Logit Regression, Decision Tree, Neural Network

Ⅰ. 서 론

1980년대 이후 각종 정보 기술의 발전으로 데이터의 수집과 축적이 매우 빠른 속도로 이루어지고 있다. 데이터의 빠른 축적은 잠재성 있는 다양한 자원들을 소유하는 것과 같은 의미를 지니지만, 또 다른 한편으로는 데이터의 대용량화와 다량화를 유발하여 데이터의 관리 및 분석을 어렵게 한다. 1990년대 후반에 들어 증가된 데이터를 효율적으로 분석하기 위한 여러 가지 방법들이 모색되었는데, 그 중에 대표적인 것이 데이터 마이닝(Data Mining)이다. 데이터 마이닝이란 일반적으로 대용량의 데이터베이스로부터 흥미롭고 유용한 패턴이나 규칙을 찾아내는 과정을 말하며, 고객관계 경영(Customer Relationship Management)에 널리 이용되고 있다(황명화, 2003).

선진국에서는 기업에서 뿐 만 아니라 정부나 각종 연구기관에서도 데이터 마이닝 기법을 활발하게 이용하고 있으나 한국에서는 그리 활성화 되지 못한 상황이었다. 그러나 최근 정부는 공공정보의 적극 공개로 국민의 알 권리 충족, 공공 빅데이터(Big Data)의 민간활용 활성화 및 데이터를 기반으로 한 과학행정 구현 등을 목표로 정부 3.0을 적극적으로 추진하고 있다. 이에 다양한 통계

* Sungkyul University (wooyang02@sungkyul.ac.kr)

정보를 포함하고 있는 공공 빅데이터를 데이터 마이닝 기법을 이용해 새로운 의미 있는 정보와 지식을 재생산할 필요성이 커지고 있는 상황이다.

이에 본 연구는 도시·부동산 분야에 데이터 마이닝 기법을 적용하고자 공공 빅데이터인 한국노동패널자료를 이용해 한국 가구 근로소득이 높은 집단과 낮은 집단 간의 정보차이를 실증분석하였다. 글로벌 금융위기 이전 년도인 한국노동패널조사 10차(2007)와 이후 년도인 15차(2012)자료를 사용해 근로소득을 종속변수로 설정하고 지역, 가구구성, 입주형태, 주택정보, 저축여부, 소유부동산 등의 변수를 독립변수로 설정하여 비교분석하였다. 데이터 마이닝 기법 중 로지스틱 회귀분석(Logit regression), 의사결정나무(Decision Tree)와 신경망 분석(Neural network)을 사용하였고 이 세 가지 분석방법의 예측력을 비교 평가하였다.

본 연구의 구성은 다음과 같다. 2장은 관련된 선행연구를 살펴보고 3장은 분석방법론으로 로지스틱 회귀분석, 의사결정나무와 신경망 분석에 대해서 구술한다. 이후 4장은 실증분석으로 각 분석모형의 결과를 구술하고 가장 예측정확도가 높은 모형을 확인한다. 마지막으로 5장은 결론으로 시사점을 제시하고자 한다.

Ⅱ. 선행연구 고찰

본 연구는 가구 소득격차와 관련된 선행연구와 도시·부동산 분야에 데이터 마이닝 기법이 적용된 선행연구로 나누어 살펴보도록 하겠다.

한국 가구 소득격차와 관련된 연구는 주로 지니계수 분해법을 이용한 연구가 주를 이루고 있는 상황이다.

정의철 외(2009)는 한국노동패널 자료를 이용하여 소득격차 수준과 그 변화를 지니계수를 이용하여 측정하고 이를 소득원천별로 지니 분해하였다. 분석결과, 근로소득이 전체 불평등에 미치는 영향력이 점차 확대되어 가고 있다고 하였으며 부동산 소득의 총소득 불평등에 대한 기여도가 증가하고 있다고 하였다.

이민환·장연주(2011)은 한국노동패널 1차년도부터 11차년도까지의 자료를 이용하여 소득원천별로 지니계수를 분해하였다. 분석결과, 1997년 가구 소득격차에 대한 상대적 기여도가 가장 낮았던 부동산소득의 기여도가 2007년에는 근로소득 다음으로 가구 소득격차에 크게 기여한다고 하였다. 또한 근로소득과 부동산소득이 높은 가구일수록 가구균등화소득이 높고, 부동산소득과 기타소득이 높은 가구일수록 가구총소득이 평균 이상으로 높다고 하였다.

최바울·김성환(2003)은 한국노동패널 1-4년차 자료를 사용해 소득분해를 이용해 분석하였다. 분석결과, 소득격차에 근로소득이 가장 큰 영향을 주고 있으나 영향력이 점차 떨어지고 있으며 고소득층의 자산소득이 소득격차에 대한 기여도가 커지고 있다고 하였다.

정진호(2001)은 1990년부터 2000년까지 도시가계조사 자료를 이용해 지니계수를 분석하였다. 분석결과, 소득격차는 경제위기 이후 급격하게 높아졌으며 소득격차에 대한 기여도는 근로소득이 가장 크게 나타났고 재산소득이 소득격차에 미치는 영향은 크지 않다고 하였다.

한국에서 도시·부동산 분야에 데이터 마이닝 기법을 적용한 사례는 그리 많지 않은 상황이다.

변루나(2001)는 2000년 도시가계조사 자료로 도시가계의 소비성향패턴을 데이터 마이닝 기법인 로지스틱 회귀분석, 의사결정나무, 신경망 분석을 이용해 분석하였다. 분석결과 로지스틱 회귀분석이 가장 예측도가 높게 나타났으며 도시가계 소비성향의 패턴에 가장 영향을 주는 변수는 총지출과

저축이라 하였다.

조주옥(2004)은 2001년도 통계청 가구소비실태조사 자료를 이용해 가구소득에 대해 데이터 마이닝 기법을 사용한 여러모형을 구축하고 가장 예측력이 좋은 모형을 찾고자 하였다. 분석결과, 의사결정나무와 신경망모형이 약간 나은 모형으로 보였지만 그 차이는 미비하다고 하였다.

김태윤·이창무(2005)는 임차인의 임대계약형태에 있어서 영향을 주는 요인을 분석하기 위해 로지스틱 회귀분석, 의사나무결정, 신경망 분석을 이용하였다. 분석결과, 세가지 분석의 예측도는 유사하게 나타났으며 각 모형이 가진 장점을 이용해 임대시장의 구조적 해석이 가능하다고 하였다.

이준용 외(2007)은 강남지역 아파트 가격을 예측하기 위해 데이터 마이닝 기법을 적용하였다. 분석결과, 데이터 마이닝 기법 중 하나만을 가지고 연구 분석기법을 사용하는 것이 아니라 각 분석방법의 장점들을 살려 연구의 질이 향상 될 수 있다고 하였다.

송호창 외(2008)는 서울시 주상복합아파트를 대상으로 주택가격과 지역 및 개발특성이 어떤 관계가 있는지를 의사결정나무를 이용해 분석하였다. 분석결과, 전용면적이라는 개발특성에 따라 일정규모 이상과 이하로 구분되었으며 이를 검증 및 유형화하여 다중회귀분석을 통해 보다 구체적이고 설명력 있는 가격영향요인을 분석하였다.

홍아름 외(2010)은 1998년부터 2009년까지의 서울시 오피스빌딩을 대상으로 부동산 간접 투자제도가 도입된 이후 서울시 오피스빌딩에 직접투자와 간접투자를 결정하는 요인을 데이터 마이닝을 이용해 실증분석하였다. 로지스틱 회귀분석, 의사결정나무분석, 신경망 분석을 사용하였고 신경망 분석이 예측의 정확도가 가장 높게 나타났다.

Ⅲ. 분석방법론

데이터 마이닝을 효과적으로 하기 위해서는 많은 사전, 사후 작업이 필요하다. 어떤 데이터가 마이닝 될 필요가 있는지 적절한 데이터를 준비하고, 마이닝에 적합한 형태로 데이터를 가공하고, 마이닝에 사용할 기법을 선택하고, 결과를 해석하는 일련의 과정이 필요하며 이러한 과정을 KDD(Knowledge Discovery in Database)라고 한다.

데이터 마이닝은 대용량(massive)의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 여기서 정보는 묵시적이고 잘 알려져 있지 않지만 잠재적으로 활용가치가 있는 정보를 의미한다. 데이터 마이닝 기법으로는 연결분석(Link analysis), 판별분석(Discrimination anaysis), 군집분석(Cluster anaysis)등의 기존 통계분석과 의사결정나무, 신경망모형, 연관성규칙(Association rule)등의 변형된 형태의 분석기법이 있다. 본 연구에서는 데이터 마이닝에서 가장 널리 이용되는 로지스틱 회귀분석, 의사결정나무, 신경망 분석을 이용하고자 한다(변루나, 2001; 배화수, 2008).

1. 로지스틱 회귀분석

로지스틱 회귀분석은 종속변수가 순서형 명목척도로 측정되어 있는 경우 종속변수와 독립변수간의 인과관계를 분석하는 방법론이다.

로지스틱 회귀분석은 종속변수가 이항형일 때 일반 선형회귀모형을 사용 못하므로 로짓 변환을 하는데 아래식과 같다.

$$\ln\frac{p(y=1|x_1,...,x_p)}{1-p(y=1|x_1,...,x_p)}=\alpha+\beta_1x_1+...+\beta_px_p \quad (1)$$

위 식(1)로부터 추정된 회귀계수를 이용하여 아래와 같이 사후 확률에 대한 추정식을 구할 수가 있다.

$$\hat{P}(y=1|x_1,...,x_p)=\frac{\exp(\hat{a}+\hat{b_1}+...\hat{b_p}x_p)}{1+\exp(\hat{a}+\hat{b_1}+...\hat{b_p}x_p)} \quad (2)$$

사후확률은 0과 1사이의 값을 가지므로 적절한 절단값을 정하여 이 값을 기준으로 개체를 분류하는 것이다. 입력변수가 분류 결정에 미치는 영향의 정도는 오즈비(Odds Ratio)로 계량화 할수 있다. 다른 모든 입력변수가 일정한 상태에서 x_i가 1단위 증가하는데 따른 오즈비는 다음과 같이 계산된다.

$$\frac{\exp(\alpha+\beta_1x_1+...+\beta_i(x_i+1)+...\beta_px_p)}{\exp(\alpha+\beta_1x_1+...+\beta_i(x_i)+...\beta_px_p)}=\exp(\beta_i) \quad (3)$$

여기서, 오즈비가 1보다 작다는 것은 입력변수 x_i가 감소방향의 영향으로 미침을 의미하고, 반대로 오즈비가 1보다 크다는 것은 증가방향으로 영향을 미친다는 것을 의미한다.

2. 의사결정나무 분석

의사결정나무는 의사결정규칙을 나무구조로 도표화하여 관심대상 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법론으로 다른 분석방법론에 비해 연구자는 분석과정을 쉽게 이해하고 설명할 수 있다.

의사결정나무 분석의 대표적인 알고리즘은 CHAID, CART 등이 있으나 본 연구에서는 CHAID를 이용하기로 한다.

CHAID(Chi-suared Automatic Interaction Detection)는 카이제곱 검정 또는 F 검정을 이용하여 다지분리(multiway spilt)를 하는 알고리즘이다. 다지분리란 부모마디에서 자식마디들이 생성될 때, 2개 이상의 분리가 일어나는 것을 허용한다는 것이다. CHAID의 종속변수가 이산형일 때, Pearson의 카이제곱 통계량 또는 우도비 카이제곱 통계량(Likelihood Ratio Chi-sqare statistic)을 분리기준으로 사용한다. 여기서 종속변수가 순서형 또는 사전 그룹화된 연속형인 경우에는 우도비 카이제곱 통계량이 사용된다. 카이제곱 통계량은 관측도수(f_{ij})로 이루어진 $r\times c$분할표로부터 계산된다.

분할표로부터, 카이제곱 통계량은 $X^2=\sum_{i,j}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$로 정의되고, 우도비 카이제곱 통계량은 $X^2=2\sum_{i,j}f_{ij}\times\ln(\frac{f_{ij}}{e_{ij}})$로 정의된다. 두 통계량의 자유도는 $(r-1)(c-1)$로 동일하다. 여기서 e_{ij}는 분포의 독립성 가설하에서 계산된 기대도수를 말하여, $e_{ij}=\frac{f_{i.}\times f_{.j}}{f_{..}}$과 같이 계산된다.

카이제곱 통계량이 자유도에 비해 매우 작다는 것은 설명변수의 각 범주에 따른 종속변수의 분포가 서로 동일하다는 것을 의미한다. 따라서 설명변수가 독립변수에 영향을 주지 않는다고 말할 수 있다.

3. 신경망 분석

신경망(Neural Network)에는 여러 가지 다양한 모형이 있으나 본 연구에서는 자료 분석시 가장

많이 이용되는 MLP(Multilayer Perceptron) 신경망을 이용하였다. MPL모형은 입력층(Input layer), 은닉마디로 구성된 은닉층(hidden layer) 그리고 출력층(output layer)으로 구성된 전방향(feed forward)신경망이다.

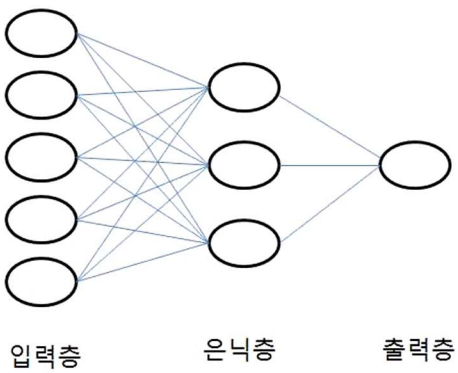


그림 1. MPL 구조
Figure 1. MPL Structure

입력층은 각 입력변수에 대응되는 마디들로 구성되어 있다. 명목형(nominal)변수에 대해서는 각 수준에 대응하는 입력마디를 가지게 되는데 이는 더미변수(dummy variable)를 사용하는 것과 같다. 은닉층은 여러개의 은닉마디로 구성되어 있다. 각 은닉마디는 입력층으로부터 전달되는 변수값들의 선형결합(linear combination)을 비선형함수(nonlinear function)로 처리하여 출력층 또는 다른 은닉층에 전달한다. 출력층은 목표변수에 대응하는 마디를 갖는다.

MPL구조를 수식으로 도식화하면 아래와 같다.

$$\begin{aligned} H_1 &= f_1(b_1 + w_{11}X_1 + b_1 + w_{21}X_2 + \ldots + b_1 + w_{p1}X_p) \\ H_2 &= f_2(b_2 + w_{12}X_1 + b_1 + w_{22}X_2 + \ldots + b_1 + w_{p2}X_p) \\ Y &= g(b_0 + w_{10}H_1 + w_{20}H_2) \end{aligned} \quad (4)$$

Ⅳ. 실증분석

본 연구에서는 글로벌 금융위기 전·후 한국 가구 근로소득이 높은 집단과 낮은 집단의 정보 차이를 분석하기 위해 한국노동연구원에서 진행한 한국노동패널조사 10차(2007년도)와 15차(2012년) 자료를 사용하였다. 글로벌 금융위기 전년도(2007년)의 자료와 입수 가능한 최근년도(2012년)자료에 대해 의사결정나무, 로지스틱 회귀분석, 신경망 분석 세 가지 데이터 마이닝 기법을 사용하여 분석하였으며 조사에 사용된 자료는 조사대상 가구의 지역, 가구구성, 입주형태, 주택정보, 경제상태, 소득구성 등 정보를 포함하고 있다. 본 연구에서는 데이터 마이닝 분석의 효율성을 고려해 이들 변수 중 근로소득과 관련성이 높은 22개의 변수만 분석변수로 선택되었다. 연구모형의 종속변수인 근로소득 집단은 글로벌 금융위기 전·후 각각 자료의 평균값을 기준으로 산출하였다. 글로벌 금융위기 전년도(2007년)의 자료의 경우 근로소득 평균값이 3268.454이므로 평균값보다 큰 경우 근로소득이 높은 집단으로 분류하였으며 평균값보다 작은 경우 근로소득이 낮은 집단으로 분류하였다.

한편 최근년도(2012년)자료의 경우 근로소득의 평균값이 3853.185이므로 평균값보다 큰 경우 근로소득이 높은 집단으로 분류하였고 평균값보다 작은 경우에는 근로소득이 낮은 집단으로 분류하였다. 연구에서 사용한 자료의 구성은 〈표 1〉과 같다.

표 1. 자료구성 Table 1. Variables

생성된 변수 Variables	설명 Description	값 Value	기준 Criteria	측정유형 Measurement
EI	근로소득 집단(평균근로소득 기준) Income Group (Average Income Basis)	1	근로소득이 높은 집단 High Income	binary
		2	근로소득이 낮은 집단 Low Income	
REGION	지역 분류 Area	1	수도권 Metropolitan	binary
		2	비수도권 Non-metropolitan	
H_TYPE	입주형태 Housing Ownership	1	자가 Own	binary
		2	기타 Not own	
HOUSE	주택종류 Housing Type	1	단독주택 Detached House	nominal
		2	아파트 Apartment	
		3	기타 Etc	
EDUC	총 교육비 Education Cost	-	-	interval
H_1410	주택 전체 평수 Housing Size	-	-	interval
H_1412	시가(만) Housing Price(10 Thousand won)	-	-	interval
H_1413	임대보증금(만) Deposit(10 Thousand won)	-	-	interval
H_1414	월세금(만) Rent(10 Thousand won)	-	-	interval
H_1501	고등학생(재수생)이하 자녀 유무 The Presence of less than High School Student	1	있다 Yes	binary
		2	없다 No	
H_1993	(고등학생 이하 자녀)경제적 부담 정도 (Less than High School Student) The Economic Burden	1	매우 부담된다 Much	ordinal
		2	약간 부담된다 Some	
		3	보통이다 Normal	
		4	별로 부담되지 않는다 Not much	
		5	전혀 부담되지 않는다 Not at all	
H_2001	대학생 이상 자녀 유무 The Presence of more than Undergraduate	1	있다 Yes	binary
		2	없다 No	
H_2061	(대학생 자녀)경제적 부담 정도 (More than Undergraduate) The Economic Burden	1	매우 부담된다 Much	ordinal
		2	약간 부담된다 Some	
		3	보통이다 Normal	
		4	별로 부담되지 않는다 Not much	
		5	전혀 부담되지 않는다 Not at all	
H_2111	금융소득 유무 The Presence of Financial Income	1	있었다 Yes	binary
		2	없었다 No	
H_2121	부동산소득 유무 The Presence of Real Estate Income	1	있었다 Yes	binary
		2	없었다 No	
H_2131	사회보험 수혜자 여부 The Presence of Social Security Beneficiaries	1	있었다 Yes	binary
		2	없었다 No	
H_2171	보호대상가구 여부 The Presence of Protected Household	1	있었다 Yes	binary
		2	없었다 No	

H_2401	작년 저축여부 The Presence of Saving (Last Year)	1	있었음 Yes	binary
		2	없었음 No	
H_2513	소유부동산 총액(범주) Total Real Estate Owned (Category)	1	1천만 원 미만 Less than 10 million won	ordinal
		2	1천-2천5백만 원 미만 Less than from 10 million won to 25 million won	
		3	2천5백-5천만 원 미만 Less than from 25 million won to 50 million won	
		4	5천-7천5백만 원 미만 Less than from 50 million won to 75 million won	
		5	7천5백만-1억 원 미만 Less than from 75 million won to 100 million won	
		6	1-2억 원 미만 Less than from 100 million won to 200 million won	
		7	2-3억 원 미만 Less than from 200 million won to 300 million won	
		8	3-4억 원 미만 Less than from 300 million won to 400 million won	
		9	4-5억 원 미만 Less than from 400 million won to 500 million won	
		10	5-10억 원 미만 Less than from 500 million won to 1 billion won	
		11	10억 원 이상 More than 10 billion	
H_2524	전세/임대보증금 총액(범주) Total Deposit(Category)	1	1천만 원 미만 Less than 10 million won	ordinal
		2	1천-2천5백만 원 미만 Less than from 10 million won to 25 million won	
		3	2천5백-5천만 원 미만 Less than from 25 million won to 50 million won	
		4	5천-7천5백만 원 미만 Less than from 50 million won to 75 million won	
		5	7천5백만-1억 원 미만 Less than from 75 million won to 100 million won	
		6	1-2억 원 미만 Less than from 100 million won to 200 million won	
		7	2-3억 원 미만 Less than from 200 million won to 300 million won	
		8	3-4억 원 미만 Less than from 300 million won to 400 million won	
		9	4-5억 원 미만 Less than from 400 million won to 500 million won	
		10	5-10억 원 미만 Less than from 500 million won to 1 billion won	
		11	10억원 이상 More than 10 billion	
H_2532	임차부동산 종류 Rental Real Estate Type	1	주택 Housing	nominal
		2	건물 Building	
		3	임야 Forests and Fields	
		4	토지 Land	
		5	기타 Etc.	
H_2652	자동차 소유 대수 Automobile ownership	-	-	interval

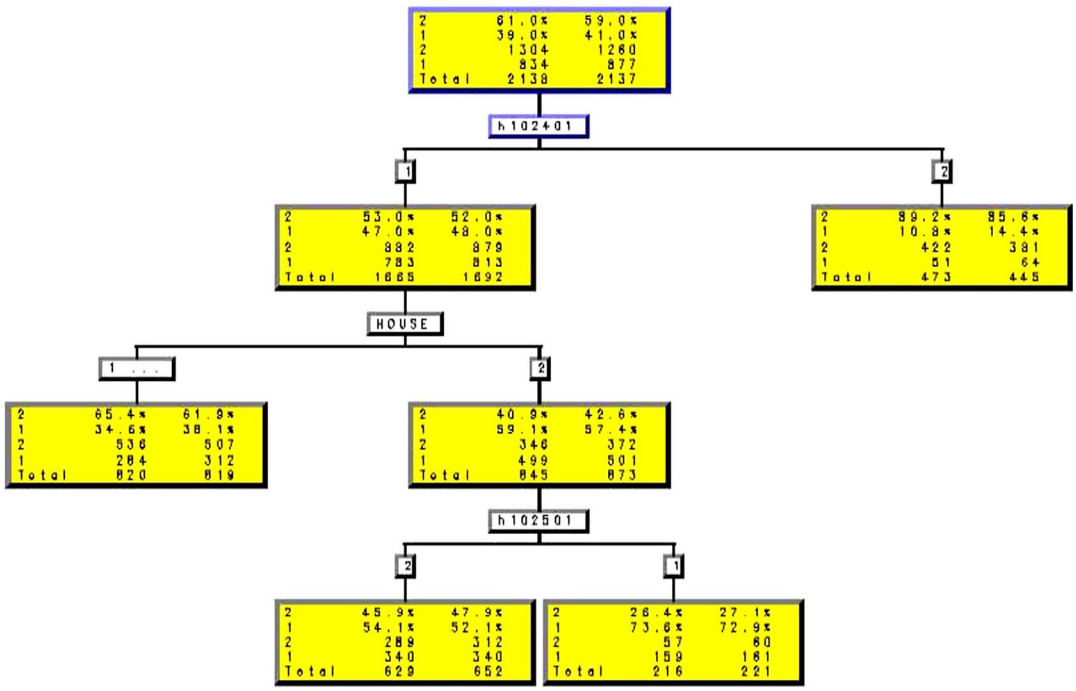


그림 3. 의사결정나무분석 결과-10차(2007년)
Figure 3. Decision Tree Analysis 10th(2007)

1. 의사결정나무 분석결과

분석에 사용한 자료는 분석용 데이터(Train Data)과 평가용(Validation Data)으로 각각 50%의 비율로 분할하였으며 앞서 선정된 변수 중에 R-square값이 0.005보다 작은 변수 또는 결측값이 50%보다 많은 변수들은 2차로 분석에서 제외하였다.

의사결정나무 분석은 E-Miner의 Decision Tree Node를 이용하였고, 분리기준(Splitting Criterion)의 알고리즘은 CHAID를 이용하였다. 의사결정나무 오분류표 결과 아래 〈표 2〉와 같다.

표 2. 의사결정나무분석 **오분류표** Table 2. Decision Tree Analysis Misclassification Table

10차(2007년) 10th(2007)				15차(2012년) 15th(2012)			
예측Prediction 실제Reality	근로소득 높음 High Income	근로소득 낮음 Low Income	총계 Total	예측Prediction 실제Reality	근로소득 높음 High Income	근로소득 낮음 Low Income	총계 Total
근로소득 높음 High Income	414	463	877	근로소득 높음 High Income	705	430	1135
근로소득 낮음 Low Income	246	1014	1260	근로소득 낮음 Low Income	366	1314	1680
총계 Total	660	1477	2137	총계 Total	1071	1744	2815
정분류율: 66.82%, Positive classification rates: 66.82%				정분류율: 71.72%, Positive classification rates: 71.72%			

10차(2007)년도의 경우 정분류율이 66.82%, 15차(2012)년도의 경우 71.72%로 15차년도의 예측력이 10차년도보다 우수함을 알 수 있다.

10차(2007)년도 근로소득이 높은 집단과 낮은 집단의 정보에 대해 의사결정나무 분석한 결과 〈그림 3〉과 같다. 근로소득이 높은 집단이 39.0%, 낮은 집단이 61.0%로 나타났다. 저축여부=있었음(h102401=1)집단의 경우 근로소득이 높은 집단이 47.0%, 낮은 집단이 53.0%로 나타났고 저축여부=없었음(h102401=2)집단에서 근로소득이 높은 집단이 10.8%, 낮은 집단이 89.2%로 나타났다.

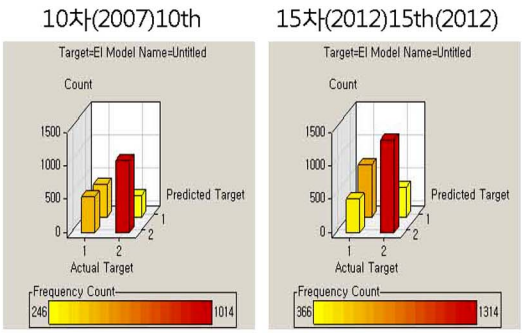


그림 2. 의사결정나무분석 모형평가
Figure 2. Decision Tree Analysis Evaluation

주택종류=아파트(house=2)집단의 경우 근로소득이 높은 집단이 59.1%, 낮은 집단이 40.9%로 나타났고 주택종류=단독주택, 기타(house=1,3)집단의 경우 근로소득이 높은 집단이 34.6%, 낮은 집단이 65.4%로 나타났다. 거주주택 외 부동산=없음(h102501=2)집단의 경우 근로소득이 높은 집단이 54.1%, 낮은 집단이 45.9%로 나타났으며 거주주택 외 부동산=있음(h102501=1)집단의 경우 근로소득이 높은 집단이 73.6%, 낮은 집단이 26.4%로 나타났다.

15차(2012)년도 근로소득이 높은 집단과 낮은 집단의 정보에 대해 의사결정나무 분석한 결과 〈그림 4〉와 같다. 근로소득이 높은 집단이 40.8%, 낮은 집단이 59.2%로 나타났다. 저축여부=있었음(h152401=1)집단의 경우 근로소득이 높은 집단이 51.1%, 낮은 집단이 48.9%로 나타났고 저축여부=없었음(h152401=2)집단에서 근로소득이 높은 집단이 16.5%, 낮은 집단이 83.5%로 나타났다.

주택종류=아파트(house=2)집단의 경우 근로소득이 높은 집단이 63.3%, 낮은 집단이 36.7%로 나타났고 주택종류=단독주택, 기타(house=1,3)집단의 경우 근로소득이 높은 집단이 37.2%, 낮은 집단이 62.8%로 나타났다. 대학생 이상 자녀 유무=없음(h152001=2)집단의 경우 근로소득이 높은 집단이 32.7%, 낮은 집단이 67.3%로 나타났으며 대학생 이상 자녀 유무=있음(h152001=1)집단의 경우 근로소득이 높은 집단이 59.5%, 낮은 집단이 40.5%로 나타났다. 입주형태=자가(h_type=1)집단의 경우 근로소득이 높은 집단이 68.7%, 낮은 집단이 31.3%로 나타났다.

입주형태=기타(h_type=2)집단의 경우 근로소득이 높은 집단이 52.0%, 낮은 집단이 48.0%로 나타났다. 결론적으로 글로벌 금융위기 전 근로소득이 높은 집단과 낮은 집단의 패턴에 가장 영향을 주는 변수는 저축(h152401)이며 그 다음으로 주택종류(house), 거주주택 외 부동산(h102501)의 순으로 구분되었다. 한편 글로벌 금융위기 후 근로소득이 높은 집단과 낮은 집단의 패턴에 가장 영향을 주는 변수도 역시 저축(h152401)으로 나타났으나, 그 다음으로는 주택종류(house), 대학생 이상 자녀 유무(h152001)와 입주형태(h_type)의 순으로 구분되었다.

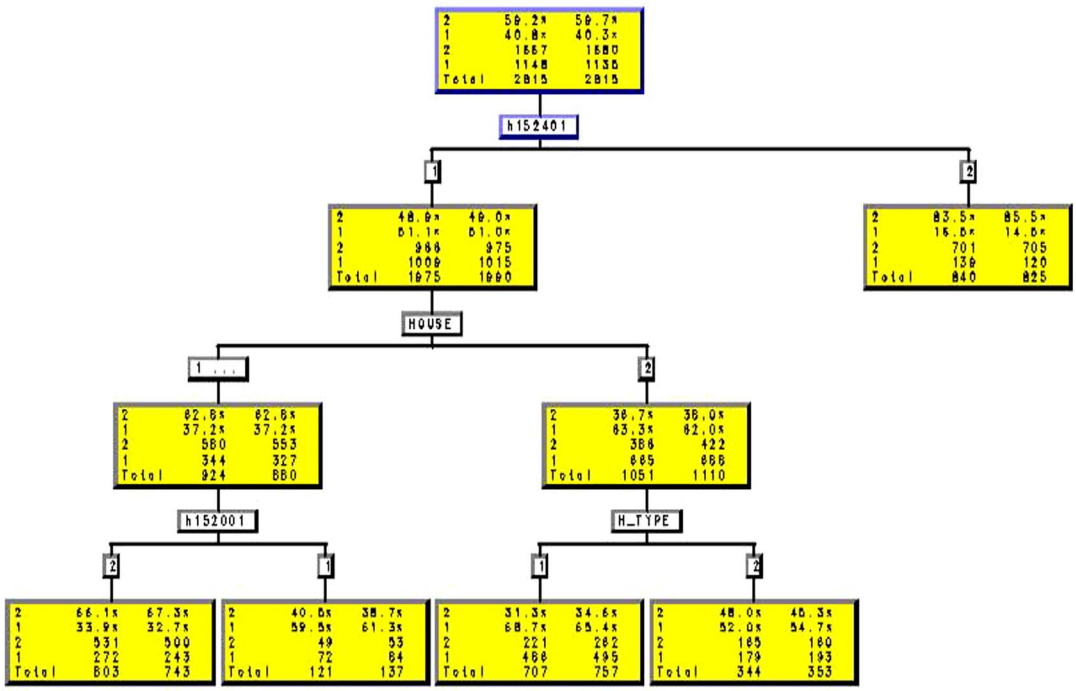


그림 4. 의사결정나무분석 결과-15차(2012년)
Figure 4. Decision Tree Analysis 15th(2012)

2. 로지스틱 회귀분석 결과

로지스틱 회귀분석에서 더미변수에 대한 코딩방식은 Deviation방식을 사용하였고, 본 연구와 관련이 있다고 선정한 모든 변수들의 영향을 살펴보기 위해 변수선택방법은 Enter(None)를 사용하였다. 평가용 자료에 대한 오분류표는 아래 〈표 3〉과 같다. 로지스틱 회귀분석 오분류표 결과 10차(2007)년도의 경우 정분류율이 69.40%, 15차(2012)년도의 경우 73.46%로 15차년도의 예측력이 10차년도보다 우수함을 알 수 있다.

표 3. 로지스틱 회귀분석 오분류표 Table 3. Logistic Regression Analysis Misclassification Table

10차(2007년) 10th(2007)				15차(2012년) 15th(2012)			
예측Prediction 실제Reality	근로소득 높음 High Income	근로소득 낮음 Low Income	총계 Total	예측Prediction 실제Reality	근로소득 높음 High Income	근로소득 낮음 Low Income	총계 Total
근로소득 높음 High Income	472	405	877	근로소득 높음 High Income	718	417	1135
근로소득 낮음 Low Income	249	1011	1260	근로소득 낮음 Low Income	330	1350	1680
총계 Total	721	1416	2137	총계 Total	1048	1767	2815
정분류율: 69.40%, Positive classification rates: 69.40%				정분류율: 73.46%, Positive classification rates: 73.46%			

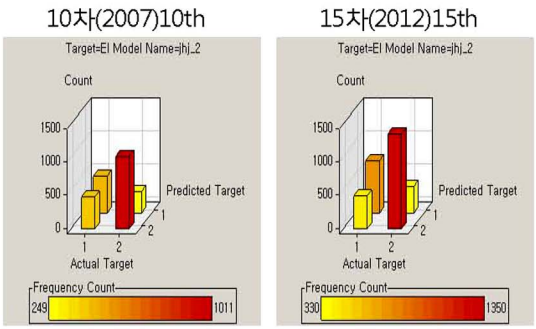


그림 5. 로지스틱 회귀분석 모형평가
Figure 5. Logistic Regression Evaluation

근로소득집단을 분석한 결과 중 로지스틱 회귀분석 결과는 〈그림 6〉과 같다. 그림에서 변수별 중요도를 의미하는 Effect T-scores는 추정한 회귀계수에서 표준오차를 나눈 값이다. 그림을 보면 왼쪽부터 오른쪽까지 Effect T-scores의 절대값 크기대로 변수들을 나열하고 있다. 우선 10차(2007)년도에서 h102401(저축여부)의 점수가 가장 높았으며 다음으로 h_type(입주형태), house(주택종류), h102001(대학생 이상 자녀 유무), h102501(거주주택 외 부동산 소유여부), h101501(고등학생 이하 자녀 유무)의 순으로 나타났다. 15차(2012)년도를 살펴보면 역시 h152401(저축여부)의 점수가 가장 높았으며 다음으로 house(주택종류), h_type(입주형태), h152001(대학생 이상 자녀 유무), h151501(고등학생 이하 자녀 유무), h152501(거주주택 외 부동산 소유여부)의 순으로 나타났다.

3. 신경망 분석 결과

신경망 분석에서는 Neural Network Node의 MLP 알고리즘을 사용하였다. 평가용 자료에 대한 오분류표는 아래 〈표 4〉와 같다. 신경망 오분류표 결과 10차(2007)년도의 경우 정분류율이 68.88%, 15차(2012)년도의 경우 73.00%로 15차년도의 예측력이 10차년도보다 우수함을 알 수 있다.

〈그림 7〉은 신경망 분석에서 평균 오차함수값의 변화를 보여준다. 그림에서는 반복 횟수에 따른 분석용 자료와 평가용 자료의 평균 오차함수(Average Error Function)를 보여주고 있으며 일반적으로 반복 횟수가 증가하면 분석용 자료에 대한 목적함수 값은 감소한다. 10차(2007)년도의 경우 평가용 자료의 평균 오차함수값이 감소하다 반복횟수 3차부터 다시 증가한 것을 확인할 수 있다. 반면 15차(2012)년도의 경우 평가용 자료는 감소하다가 반복횟수 31차부터 다시 증가한 것으

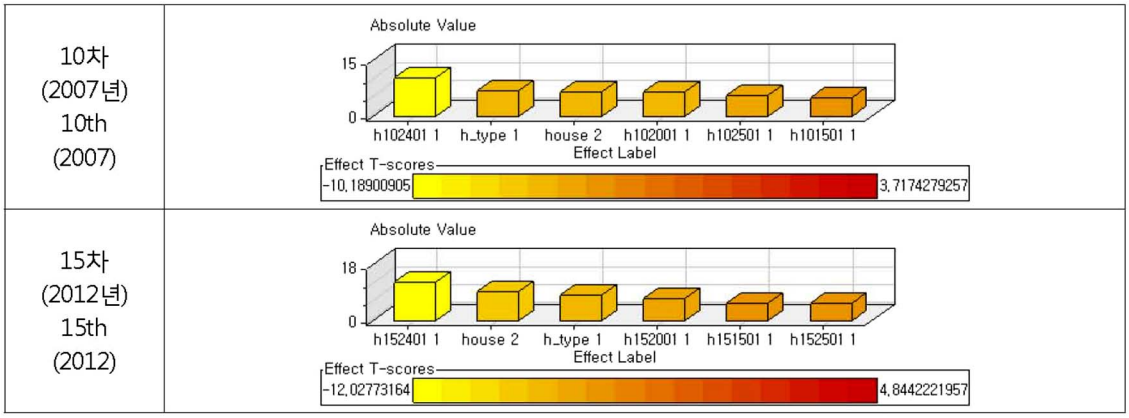


그림 6. 로지스틱 회귀분석 결과
Figure 6. Logistic regression analysis

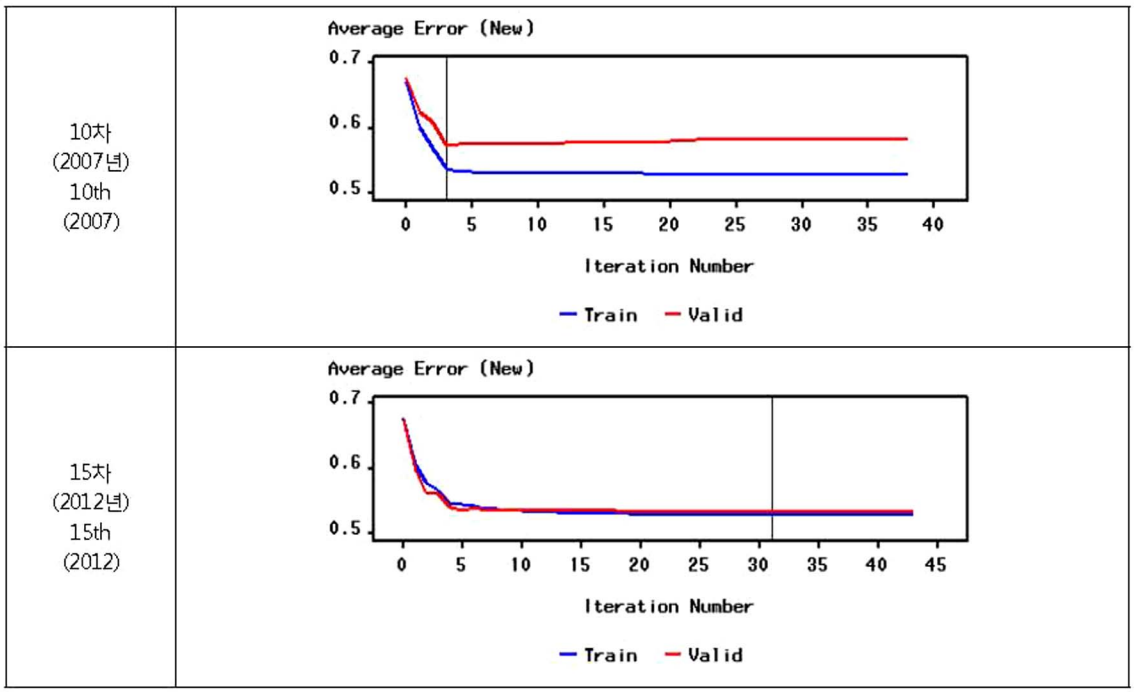


그림 7. 신경망분석 평균 오차함수값의 변화

Figure 7. The Change of average value of the error function of the neural network analysis

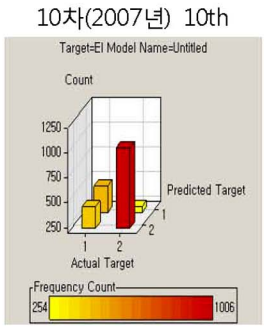
로 나타났으나 분석용 자료와 거의 차이가 없는 것을 확인할 수 있다.

4. 모형평가 및 최종모형 선택

앞서 세 가지 분석 결과를 종합해 보면 아래 〈표 5〉와 같다. 결과를 보면 10차 년도와 15차 년도에서 모두 로지스틱 회귀분석의 정분류율이 가장 우수함을 알 수 있다.

표 4. 신경망분석 오분류표 Table 4. Neural Network Analysis Misclassification Table

10차(2007년) 10th(2007)				15차(2012년) 15th(2012)			
예측Prediction 실제Reality	근로소득 높음 High Income	근로소득 낮음 Low Income	총계 Total	예측Prediction 실제Reality	근로소득 높음 High Income	근로소득 낮음 Low Income	총계 Total
근로소득 높음 High Income	466	411	877	근로소득 높음 High Income	770	365	1135
근로소득 낮음 Low Income	254	1006	1260	근로소득 낮음 Low Income	395	1285	1680
총계 Total	720	1417	2137	총계 Total	1165	1650	2815
정분류율: 68.88%, Positive classification rates: 68.88%				정분류율: 73.00%, Positive classification rates: 73.00%			



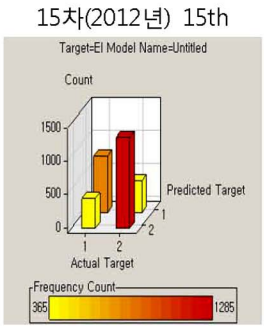
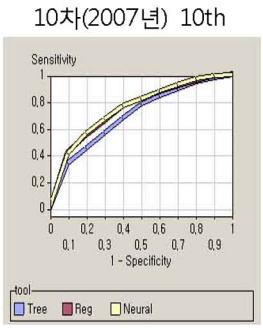


그림 8. 신경망분석 모형평가
Figure 8. Neural Network Analysis Evaluation



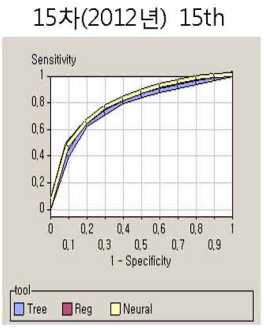


그림 9. ROC곡선
Figure 9. ROC Curve

한편, 세 가지 모형의 성능을 민감도(Sensitivity)/특이도(Specificity) 기준으로 그려진 ROC곡선을 출력한 결과 〈그림 9〉와 같다. 그림을 보면 세 가지 모형 모두 우상향하여 모형의 구축 효과는 입증되고 있으며 로지스틱 회귀모형과 신경망모형은 거의 차이가 없는 것을 확인할 수 있다.

표 5. 모형평가 Table 5. Methodology Evaluation

분석방법 Methodology	10차(2007년) 10th(2007)	15차(2012년) 15th(2012)
의사결정나무 Decision Tree	66.82%	71.72%
로지스틱 회귀분석 Logistic Regression	69.40%	73.46%
신경망분석 Neural Network	68.88%	73.00%

V. 결 론

본 연구에서는 글로벌 금융위기 전·후 한국 가구 근로소득이 높은 집단과 낮은 집단의 정보 차이를 분석하기 위해 한국노동연구원에서 진행한 한국노동패널조사 10차(2007년도)와 15차(2012년) 자료를 사용하였다. 글로벌 금융위기 전년도(2007년)의 자료와 입수 가능한 최근년도(2012년)자료에 대해 의사결정나무, 로지스틱회귀분석, 신경망분석 세 가지 데이터 마이닝 기법을 사용하여 분석하였으며 조사에 사용된 자료는 조사대상 가구의 지역, 가구구성, 입주형태, 주택정보, 경제상태, 소득구성 등 정보를 포함하고 있다. 본 연구에서는 데이터 마이닝 분석의 효율성을 고려해 이들 변수 중 근로소득과 관련성이 높은 22개의 변수만 분석변수로 선택되었다.

의사결정나무, 로지스틱 회귀분석, 신경망 분석에 의해 산출된 모형의 예측정확도는 대체로 만족할 만큼 높게 나타났다. 그 중에서 로지스틱 회귀분석의 예측도가 가장 높게 나타났다.

로지스틱 회귀분석결과, 글로벌 금융위기 이전 시기인 10차(2007)년도에서 저축여부의 점수가 가장 높았으며 다음으로 입주형태, 주택종류, 대학생

이상 자녀 유무, 거주주택 외 부동산 소유여부, 고등학생 이하 자녀 유무의 순으로 나타났다. 글로벌 금융위기 이후시기인 15차(2012)년도를 살펴보면 역시 저축여부의 점수가 가장 높았으며 다음으로 주택종류, 입주형태, 대학생 이상 자녀 유무, 고등학생 이하 자녀 유무, 거주주택 외 부동산 소유여부의 순으로 나타났다. 한국 가구 근로소득이 높은 집단과 낮은 집단의 차이에 저축이 가장 큰 영향을 미치는 것으로 나타났으며 그 다음으로는 주택소유 및 입주형태여부가 영향을 미치는 것으로 나타났다. 글로벌 금융위기 전·후의 차이점을 살펴보면 이후기간에는 글로벌 금융위기로 인해 부동산가격이 하락하여 거주주택 외 부동산 소유여부가 근로소득 집단 차이에 미치는 영향력이 줄어든 것을 살펴볼 수 가 있다.

본 연구는 도시·부동산 분야의 공공 데이터에 데이터 마이닝 기법을 적용해 모형을 구축함에 의의가 있으며 더욱 다양한 변수를 이용해 연구를 확장하는 것은 추후 연구과제로 남긴다.

인용문헌
References

1. 김태윤·이창무, 2005, “임차인의 임대계약 선택에 있어서 데이터 마이닝기법들을 이용한 비교분석”, 대한국토·도시계획학회 정기학술대회, 서울: 중앙대학교
   Kim, T. Y. & Lee, C. M. , 2005, "Comparative Study on Renter's Choice with Data Mining Techniques", *Korean Planners Association Congress,* Seoul: Chungang University
2. 변루나, 2001, “데이터마이닝 기법을 이용한 도시가계 소비성향 분석”, 「통계분석연구」, 6(2):85-111
   Byon, L. N., 2001, "Analysis of the Propensity to Consume for Urban Household Using Data Mining Technique", *Journal of the Korean Official Statistics*, 6(2):85-111
3. 배화수, 2008, SAS Enterprise Miner를 이용한 데이터마이닝, 서울: 교우사
   Bae, H. S., 2008, *SAS Enterprise Miner Data Mining*, Seoul: Kyowoo
4. 송호창·김태호·이주형, 2008, “주상복합아파트의 주택규모별 가격결정요인 분석” 「서울도시연구」, 9(3):79-92
   Song, H. C., Kim, T. H., Lee, J. H., 2008, "An Analysis of the Price Determinant Factors of Mixed-use Development by Housing Scale", *Seoul Studies*, 9(3):79-92
5. 이민환·장연주, 2011. “한국 지역별 가구소득 격차와 결정요인에 관한 실증분석”. 「산업혁신연구」, 27(3):111-138
   Lee, M. H.&Jang, Y. J., 2011, "An Empirical Study on the Household Income Inequality and Its Determinants in Korea", *Journal of Business and Economics*, 27(3):111-138
6. 이준용·최미화·이상엽, 2007, “데이터 마이닝 적용을 통한 아파트 가격예측에 관한 연구”, 「국토계획」, 42(4):135-148
   Lee, J. Y., Choi, M. H., Lww, S. Y., 2007, "A Study on the Forecasting Model of Apartment Price Based on Data Mining", *Journal of Korean Planners Association*, 42(4):135-148
7. 정의철·김진욱·하두나, 2009. “부동산소득이 소득불평등에 미치는 영향 분석”. 「주택연구」, 17(2):5-28
   Chung, E. C., Kim, G. U., Ha, D. N., 2009, "An Analysis on Effect of Property Income on Income Inequality", *Housing Studies Review*, 17(2):5-28
8. 정진호, 2001, “최근의 소득불평등도 변화와 소득원천별 분해,” 「노동정책연구」, 1:1-18
   Jung, J. H., 2001, "Recent Inequality

Changes and Income Source Decomposition", *Journal of Labor Policy*, 1:1-18

9. 조주옥, 2004, "데이터 마이닝 기법을 이용한 한국 가구소득에 대한 모형 연구", 단국대학교 석사학위논문
   Cho, J. O., 2004, "A Study on the Model of the Korean Household Income Using Data Mining", Master's Degree Dissertation, Dankook University.
10. 최바울·김성환, 2003, "경제위기와 소득 불평등:1997년 이후를 중심으로", 제4회 노동패널 학술대회 자료집, 개최지역: 개최장소
    Choi, B. W. & Kim, S. H., 2003, "Economic Crisis and Inequality", The 4th Korea Labor Panel Congress
11. 황명화, 2003, "공간 데이터 마이닝 방법론에 대한 연구:서울시 노령인구 분호를 사례로", 서울대학교 석사학위논문
    Hwang, M. H., 2003, "A Study on the Methodology of Spatial Data Mining:The Case Study on the Spatial Distribution of Elderly Population in Seoul", Master's Degree Dissertation, Seoul National University.
12. 홍아름·고재풍·유선종, 2010, "데이터 마이닝을 이용한 서울시 오피스빌딩 투자특성에 예측에 관한 연구", 「서울도시연구」, 11(2):51-68
    Hong, A. R., Ko, J. P., Yoo, S. J., 2010, "A Study on the Forecasting Model of the Investment Characteristics of Seoul Office Buildings based on Data Mining", *Seoul Studies*, 11(2):51-68
13. http://www.kli.re.kr/

Date Received 2014-12-01
Date Reviewed 2015-01-21
Date Accepted 2015-01-21
Date Revised 2015-02-01
Final Received 2015-02-01