응용통계연구 (2009)
22(1), 41–57

세계 각국의 자원에 대한 통계적 고찰

허문열[1] · 최병수[2] · 이승천[3]

[1]성균관대학교 통계학과, [2]한성대학교 멀티미디어학과, [3]한신대학교 정보통계학과

(2008년 11월 접수, 2008년 12월 채택)

요 약

본 논문에서는 세계 232 개국에 대한 인구, 경제 및 기타 자원에 관한 자료를 사용하여 국가의 개발정도, 인간개발지수, 경제력 그리고 OECD가입 여부에 어떤 자원이 어떻게 영향을 미치는가를 통계적으로 고찰해보고자 한다. 여기서 사용하는 국가별 자원 자료는 연속형 자료와 이산형 자료가 혼재되어있는 혼합형이며 많은 결측값이 포함되어 있어 기존의 방법으로는 분석하는 데 한계가 있다. 이 논문에서는 시각적 방법을 동원하여 복합형 자료를 탐색하는 과정을 제시하고 이러한 방법의 한계점을 보이고자한다. 이러한 한계점을 극복하고 객관적인 판단기준을 적용하여 주어진 문제에 대한 과학적인 결론을 유도하기위해 Shannon (1948)의 엔트로피 이론에 기본을 둔 상호정보(MI)를 활용하고자 한다. 상호정보를 추정하는 방법은 여러 가지가 있으며 각 방법에 따라 결과가 매우 다르게 나타난다. 본 논문에서는 Fayyad와 Irani (1992)의 이산화 방법을 적용하여 MI를 추정하는 방법을 적용한다. 여기서 이루어지는 모든 과정은 다차원 자료의 시각적 탐색 도구인 DAVIS (Huh와 Song, 2002)를 사용하였다.

주요용어: 평행좌표계, 평행 상자도형, 데이터의 시각화, 상호정보, DAVIS.

1. 서론

세계 각국은 이제 국경이 없어지고 보이지 않는 경쟁 속에서 서로 긴밀한 관계를 유지하면서 공존해가고 있다. 이러한 환경에서 대한민국은 세계에서 어떤 위치에 있는가? 또, 이웃하는 일본과는 어떤 면에서 차이가 있고, 향후 중국의 국제적인 지위가 어떻게 될 것인가? 여기에 대해 많은 학자들, 특히 역사학자, 정치학자, 경제학자들이 그들 나름대로의 이론에 의해 예측을 하고 있다. 통계학자들의 주장은 통계를 보면 세상이 보인다고 한다. 이는 통계적인 자료를 분석함으로서 그 집단의 정보를 알 수 있다는 뜻이다.

본 논문에서는 세계 각국의 자원을 조사하고, 어떤 자원을 갖는 나라가 경제적으로 부유하고, 높은 국민복지를 갖고 있으며, 발전된 나라인가를 통계적으로 살펴보고자 한다. 이를 위해 세계 232개국의 40여 개 자원을 수집하였다. 자료는 미국 CIA 자료집, OECD 보고서, UNICEF, World Bank, 각 나라의 홈페이지 등 여러 곳에서 발췌하였다. 이 변수들은 크게 인구관련, 경제관련 그리고 기타로 나누어진다. 이들 자료는 가능하면 가장 최근에 발표된 자료를 수집하려고 하였으나 몇 개의 자료에서는 이것이 가능하지 못하였다. 따라서, 어떤 자료는 2000년에 작성된 것이고 어떤 자료는 2008년에 작성된 것이다. 또한 자료에 따라 결측값이 매우 많은 것들이 있다. 예를 들어 IQ는 전체 232개국 중 80개국만 자료가 있었다.

본 연구는 2007년도 한성대학교 교내연구비 지원과제임.
[3]교신저자: (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수. E-mail: seung@hs.ac.kr

본 논문에서 가장 중요한 내용이 기초 자료이고, 어떤 기준으로 이들 변수를 선택하였느냐 하는 문제가 발생할 수 있다. 또, 자료의 작성 기관과 시점이 다르기 때문에 이를 일괄적으로 고려하여 분석하는 것에 대해 문제를 제기할 수 있다. 또한 자료의 정확성이 문제가 될 수 있다. 즉, 어떤 기관에서 자료를 작성하였는가에 따라 자료 자체의 값이 차이가 날 수 있다. 여기서는 이러한 모든 문제에 대해 일단 문제가 없다고 가정한다. 즉, 주어진 자료가 일관성이 있어 서로 비교 가능하다고 가정한다.

자료의 형태는 연속형과 이산형이 섞여 있으며 대부분의 자료에서 매우 높은 비율의 이상값들이 나타나고 있다. 예를 들어 각국의 인구를 살펴보면 중국, 인도와 같이 매우 큰 나라가 있는가 하면, 대부분 나라들은 인구가 1,000만 이하이다. 따라서 이들의 평균과 표준편차는 별 의미를 갖지 못한다. 이러한 자료로부터 정보를 얻기 위해서는 먼저 자료의 탐색이 이루어져야 한다. 여기서 정보는 이미 언급한 바와 같이 국가의 경제력이나 웰빙지수 등에 국가의 다른 자원들이 어떤 영향을 미치는가 이다. 즉, 본 논문에서 추구하는 정보를 위한 통계적 방법은 혼합변수들 간의 연관성 추정이다. 이를 알아보기 위해서 본 논문에서는 두 가지 방법을 사용하고 있다. 첫째는 데이터 시각화 방법이다. 데이터 시각화 방법에서 가장 많이 활용되는 방법인 평행좌표계와 평행상자도형을 사용하면 우리가 관심이 있는 어떤 집단이 전체에서 어떤 수준에 있는가를 용이하게 알아 볼 수 있다. 예를 들어, OECD 가입국과 비가입국은 어떤 자원에서 차이가 있는가? 또는, 한국과 일본은 어떤 면에서 차이가 있는가? 등을 쉽게 알아 볼 수가 있다. 그러나 데이터 시각화 방법은 다변량 복합자료에서 매우 효용성이 높은 방법이지만, 적용하는 사람에 따라 방법이 다를 수 가 있으며 결과의 해석이 다를 수 있다. 따라서, 두 번째 방법으로 상호정보에 의한 연관성 척도를 고려하고자 한다. 이 방법은 Shannon (1948)의 엔트로피 이론에 기반을 두고 있다. 이 방법은 매우 일반적이고 어떤 형태의 변수에도 적용이 가능하다. 그러나 이를 추정하는데 문제가 있어 사용에 많은 제한이 있었다. 본 논문에서는 Fayyad와 Irani (1992)의 MDL(minimum description length)원리에 의한 이산화 방법을 적용하여 상호정보를 추정하고 이를 통해 본 논문의 핵심인 국가별 경제력, 웰빙지수 등에 영향을 가장 많이 미치는 자원들을 찾아내고자 한다. 또한 이 결과를 사용하여 한국의 취약점과 강점을 찾아내고자 한다.

2. 자료에 대한 설명

본 논문에서 사용하는 자료는 세계 232개국에 대해 42개의 인구 경제학적 자료로서 미국 인구센서스국, 미국 CIA, IMF, UN 등과 각 나라의 홈페이지에서 발췌하였다. 232개국의 나라와 각 변수들은 다음과 같다.

1. 국가: 국가 명(Country)
2. 국토, 인구관련: 국토면적(Area, km^2), 내륙여부(Ocean), 5개 대륙(Region), 인구중위수(Pop05), 65세 이상 인구비율(Pop65), 출생률(BirthR), 사망률(DeathR), 유아 사망률(InfMor), 기대수명(LifeEx), 여성 가임률(FertR), 성비(SexR), AIDS 감염률(AIDSR), 도시인구율(PopUrb),
3. 경제 관련: GNI, GDP(per capita), 에너지 소비량(Energy), 외채(DebtExt), 노동력(KabF), 노동력 비율(LabF2), 이주율(MigRatio), 빈민인구비율(PovertPop), 실업률(UnempR), GINI, GINI 10% 상/하 비율(GINI_10), 경제력(IncomeGroup, GDP를 4개 구간으로 구분), 국가 개발등급(Developm, 3개 등급), 남자나이 중위수(AgeM), 여자나이 중위수(AgeW), 나이 중위수(AgeT)
4. 기타: 생활만족도(LifSat), 문자해독률(Literacy), 보건지출비용(HealthEX), 올림픽 메달 수(OlympicM), 영어가능자 비율(EnglishSpeakers), 인터넷 보급률(Internet), 자동차 보유대수 비

표 **2.1.** 연속형변수의 기술통계량

변수명	평균	표준편차	최소값	Q25	중위수	Q0.75	최대값	결측치
Area	572,076	1,712,776	0.440	6,020	88,361	447,400	17,098,242	12
IQ	88.500	11.766	59.000	81.000	90	98	107	152
Pop05	29,390	121,152	0.050	408.000	4,839	18,549	1,326,910	8
Pop65	7.275	4.869	1.000	3.000	5.	11	22	10
AgeMed	26.975	8.309	15.000	19.200	25.3	34.6	42.9	68
BirthR	22.240	11.204	7.600	12.500	20.1	28.1	49.6	41
DeathR	9.038	4.231	1.400	5.900	8	11.2	22.1	38
InfMor	32.871	33.463	2.300	7.700	19.51	51.920	182.31	23
LifeEx	68.250	11.697	32.230	62.840	72.42	76.880	83.52	23
PopUrb	56.269	25.451	0.000	36.000	56	78	100	5
FertR	2.881	1.538	1.000	1.750	2.39	3.71	7.340	22
SexR	1.004	0.112	0.770	0.960	0.99	1.02	1.870	26
Migratio	−0.296	3.626	−12.590	−0.870	0.00	0.66	15.660	68
AIDSR	2.214	4.941	0.010	0.100	0.30	1.09	28.090	66
GDP	12,248	19,350	56.000	976	3,857	13,630	106,082	50
GNI	11,022	11,397	596.000	2,284	6,658	16,627	66,821	40
HDI	0.728	0.171	0.336	0.583	0.773	0.863	0.968	54
GINI	40.389	10.138	23.000	33	39	46.9	70.70	101
GINI_10	20.450	22.737	4.300	8.6	12.3	21.6	157.3	107
Energy	2,666	3,046	160.900	640.	1,545	3,761	21,396	103
DebtExt	56.480	158.15	0.000	10.19	21.240	41.28	1,844	33
LabF	14,612	67,205	0.015	156.7	2,262	7,400	803,300	20
LabF2	41.833	13.737	0.490	33.85	42.53	49.23	133	23
UnempR	12.750	14.715	0.000	4.4	7.6	15	90	46
PovertyP	32.099	19.424	0.950	17	30	45	86	93
LifeSat	5.996	1.162	3.000	5.1	6.1	7	8.2	68
Literacy	82.586	19.656	23.600	71.4	90.4	98.8	100	59
HealthEx	713.282	988.361	15.000	96	309	745	5,711	76
OlympicM	103.846	269.571	1.000	2	14	72	2,514	109
EnglishS	41.851	33.741	0.000	10.36	37.93	79.38	100	110
Vehicles	176.117	193.448	0.001	18	98	289	765	92
Internet	40.710	25.363	3.700	19	41.6	63.6	90.1	161

율(Vehicles), 직업(Occup, 농업, 공업, 서비스업), 종교(Religion), 인간개발지수(HDI), 인간개발지수등급(HDIC, 3개 등급으로 구분), IQ, OECD 가입여부(OECD)

여기서 Ocean, Region, IncomeGroup, Developtment, Occupation, Religion, HDIC, OECD 8개는 이산형이고, 그 외의 다른 변수는 모두 연속형이다. 또, Occupation은 농업, 공업, 서비스업 중 각 나라의 인구가 가장 많이 차지하고 있는 부분을 그리고 Religion은 그 나라의 종교 중 가장 많이 차지하고 있는 것을 나타낸다.

자료의 전체적인 흐름을 살펴보기 위해 연속형 자료는 평균, 표준편차 등 5의 통계량을, 이산형 자료의 경우 각 범주별로 빈도수를 구한 것이 표 2.1과 2.2에 나타나 있다. 각 변수에 대한 보다 자세한 설명과 자료의 출처는 `http://www.stat.skku.ac.kr/myhuh/data/world.html`를 참조할 수 있다.

표 **2.2.** 범주형변수의 범주 및 관찰돗수

변수명	범주(돗수)
Region	Africa(56), America(33), Asia(50), Carib(17), Europe(43), Oceania(17), 결측치(26)
Developm	less_developed(38), developing(125), developed(36), 결측치(33)
OECD	N(202), Y(30), 결측치(0)
HDIC	L(22), M(85), H(71), 결측치(54)
incomeGr	L(49), LM(53), UM(40), H(67), 결측치(23)
OCEAN	coast(129), inland(35), 결측치(69)
Religion	else(19), Buddhism(9), Christianity(93), Confucianism(1), Hinduism(3), Islam(40), 결측치(67)
Occup	agriculture(16), industry(25), service(153), 결측치(38)

3. 시각적 탐색 방법에 의한 각국 자원의 비교 분석

국가별 자원에 관한 자료에 숨겨져 있는 정보를 탐색하기 위해 시각적 탐색방법을 사용한다. 여기서 사용하는 시각적 탐색방법은 동적그래픽스 기법에 기반을 두고 있으며, 이들은 모두 시각적 탐색도구인 DAVIS (Huh와 Song, 2002)로 구현되었다.

시각적 자료탐색에서 많이 활용되고 있는 도형은 산점도 행렬, 평행상자도형, 평행좌표계, FEDF 등이 있다. 그러나 산점도 행렬은 자료의 차원이 10개가 넘으면 표현하는 데 한계가 있으며, 평행상자도형도 연속형 자료에만 사용이 가능하다는 문제가 있다. 이에 반해 평행좌표계와 FEDF는 이산형 자료와 연속형 자료를 동시에 표현할 수 있다는 장점이 있다. 이 두 가지 방법은 특히 다변량 복합자료를 표현하는 데 매우 효율적이며 동적그래픽스를 구현하는 데 적절한 도구로 사용할 수 있다.

3.1. 한국의 수준은 전 세계 국가에서 어느 정도인가?

그림 3.1의 평행좌표계에서 한국은 붉은 색으로 표시되어 진하게 나타내었다. 또, 평행상자도형에서는 따로 하나의 점으로 나타나 있는 것이 한국이다. 전반적으로 한국의 지위를 살펴보면 한국은 OECD 가입국이며, 직업에서 서비스업으로 분류된다. 이밖에 인간개발지수는 높은 반면, 출산률, 유아 사망률, 여성 가임률, AIDS 환자비율, GINI_10%비율, 실업률, 외채비율 등이 상대적으로 매우 낮게 나타나고 있다. 실제로 여성 가임률은 홍콩, 마카오 다음으로 세계에서 가장 낮은 나라이다. 한편, 기대수명, IQ, 문자 해독률 등은 매우 높은 것으로 나타나 있다. 상자도형을 참고하여도 이러한 내용을 살펴볼 수 있지만, 평행좌표계에 비해 이해하는데 더 어려운 것을 알 수 있다. 이는 이미 잘 알려진 바와 같이 직선으로 이루어지는 도형이 시각적 정보를 제공하는데 더 효율적인 것을 확인해 주는 내용이다.

그림 3.1의 평행좌표계에서는 남녀 나이의 중앙값에서 남자의 경우가 여자의 경우보다 더 큰 것처럼 보인다. 그러나 이것은 남자는 14.9, 여자가 51.9로 매우 특이한 값을 갖는 Uganda 때문에 나타난 것이다. 이 나라를 제외하고 다시 평행좌표계를 그리면 한국인의 경우 여자 수명의 중앙값이 남자에 비해 약간 크다는 것을 알 수 있다.

이제 한국과 일본 두 나라를 비교해 보자. 즉, 한국과 일본은 국가 자원에서 어떤 점이 비슷하고 어떤 점에서 차이가 있는가를 살펴보기로 하자. 그림 3.2의 평행좌표계에서 붉은색은 한국이고 남색은 일본을 나타내고 있는 바, 이를 살펴보면 한국과 일본 두 나라간의 비교가 한눈에 들어온다. 즉, 이 그림에서 이산형 특성을 나타내는 좌측의 9개 좌표 중에서 종교만 제외하면 모두 비슷한 성격을 갖는다. 즉, Religion에서 한국은 불교, 일본은 기타로 분류되고 있으나, 그밖에 이산형 특성은 매우 비슷하였다. 한편, 인구 경제적 특성과 기타 자원을 비교하면, 한국의 사망률이 일본보다 작고, GINI 계수가 한국이 일

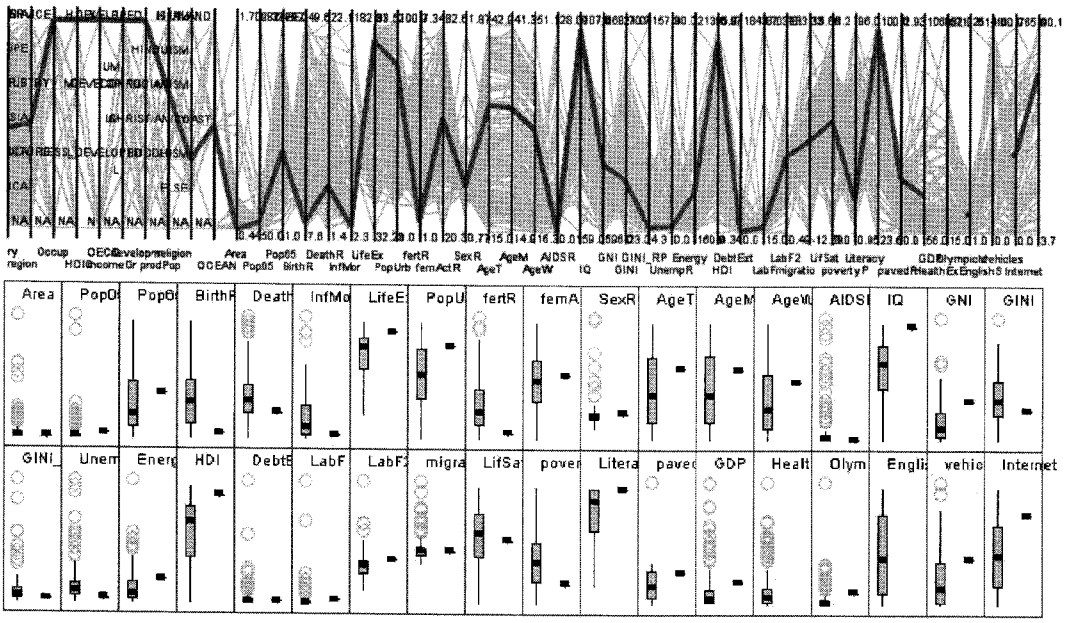

그림 **3.1.** 한국의 위치를 표시한 평행좌표계와 평행 상자도형

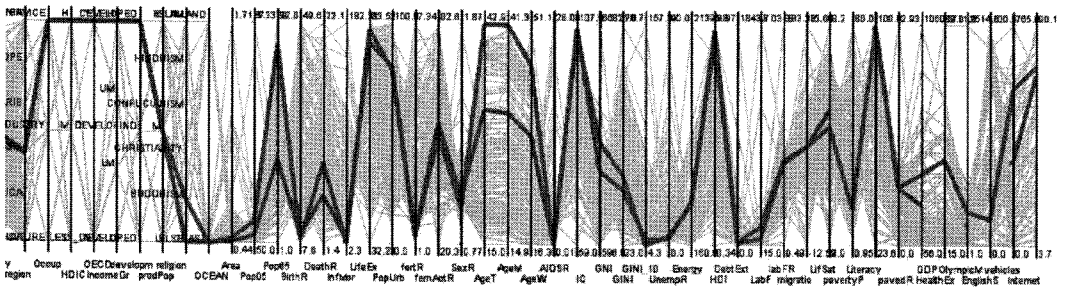

그림 **3.2.** 한국과 일본의 비교를 위한 평행좌표계(붉은색: 한국, 남색: 일본)

본에 비해 작은 반면, GINI_10은 한국이 더 높아, 일본의 소득 균형이 더 잘 되어 있었다. 또, 일본은 노동력과 생활 만족도가 한국보다 약간 높았으며, GNI, GDP 등 경제력과 올림픽 메달 수, 인구 1,000명 당 차량 수에서도 한국보다 높았다.

중국과 인도의 비교를 위한 평행좌표계는 그림 3.3과 같다. 그림에서 붉은색은 중국, 남색은 인도를 나타낸다. 이를 살펴보면 중국은 공업국가인 반면, 인도는 서비스산업 중심 국가이다. 중국은 65세 이상 인구비율이 더 많고, 출생률과 여성 가임률은 작으며, GINI 계수가 더 높고, 인간개발지수, 생활만족도, 노동력, 문자 해독률, 올림픽 메달 수, 인터넷 보급률 등 모든 부분에서 인도보다 더 높은 것으로 나타났다. 즉, 중국은 국가 자원에서 인도에 비해 전반적으로 매우 우월하다는 것을 알 수 있다.

3.2. OECD 가입국은 비 가입국에 비해 어떤 특성이 있는가?

그림 3.4에는 OECD 가입국과 비가입국을 붉은색과 남색으로 구별하여 평행좌표계와 평행상자도형을

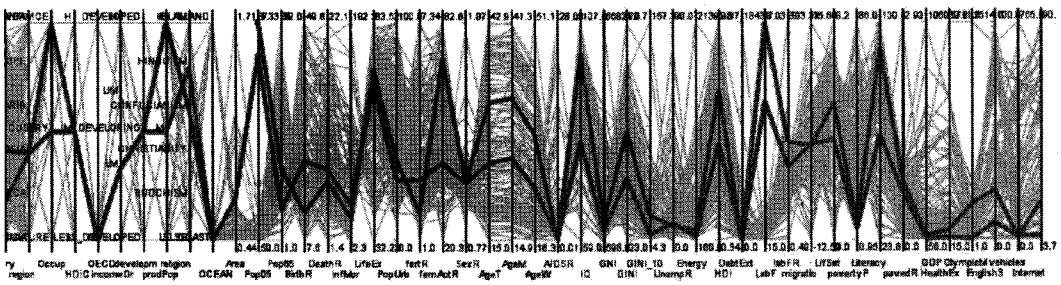

그림 **3.3.** 중국과 인도의 비교를 위한 평행좌표계(붉은색: 중국, 남색: 인도)

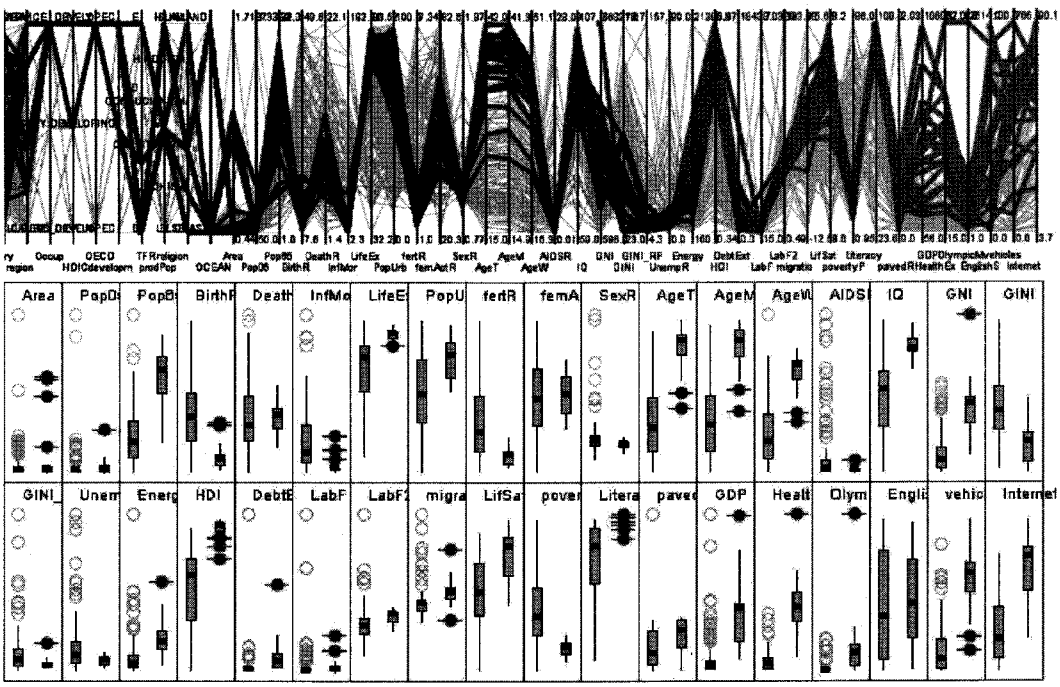

그림 **3.4.** OECD 가입국의 특징을 파악하기 위한 평행좌표계와 평행 상자도형(붉은색: 가입국, 남색: 비가입국)

그린 것이 나타나 있다. 이 내용을 살펴보면 가입국과 비가입국 간에 자원의 차이가 어떻게 나타나는가를 알 수 있다. 전반적인 흐름을 보면 OECD가입국은 서비스업이 중심이고, 한국을 제외하면 불교국가는 없었으며, 아프리카에는 OECD 가입국이 없었고, 출생률, 유아 사망률, 여성 가임률, 성비, AIDS 감염률 등이 낮은 반면, 65세 이상 인구비율과 인구연령 중앙값, 보건지출비용, IQ, 생활만족도, 문자해독률 등이 높았다. 당연히 경제력을 측정하는 변수(GNI, GDP, HDI, energy)에서도 높은 값을 보이고 있다. 이런 내용에 반해 사망률, 성비, 노동력, 영어회화 가능자 비율 등은 별 차이가 없는 것으로 나타났다. 그러나 각 변수들이 많은 이상값을 가지고 있어 이들을 시각적 방법으로만 판단하는 것은 매우 위험하다.

예를 들어, 그림 3.5와 같이 OECD 가입여부, 외채, GDP, 3개의 변수에 대해 평행좌표계를 그리고 두 개의 연속형 변수인 외채와 GDP를 OECD 가입 여부에 따라 구분하여 상자도형을 표시하면, 비가

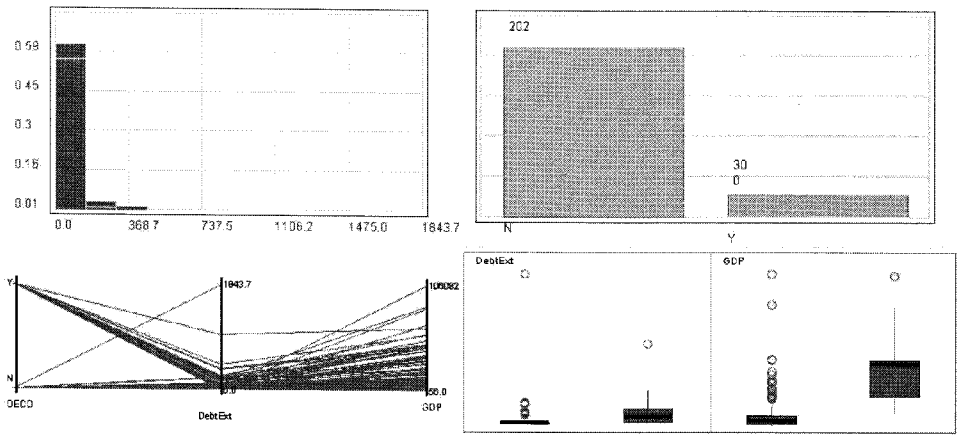

그림 3.5. OECD 가입국을 붉은색으로 표시. 비가입국 중 하나인 MONACO와 가입국 중 IRELAND의 외채가 다른 국가들에 비해 매우 높음.

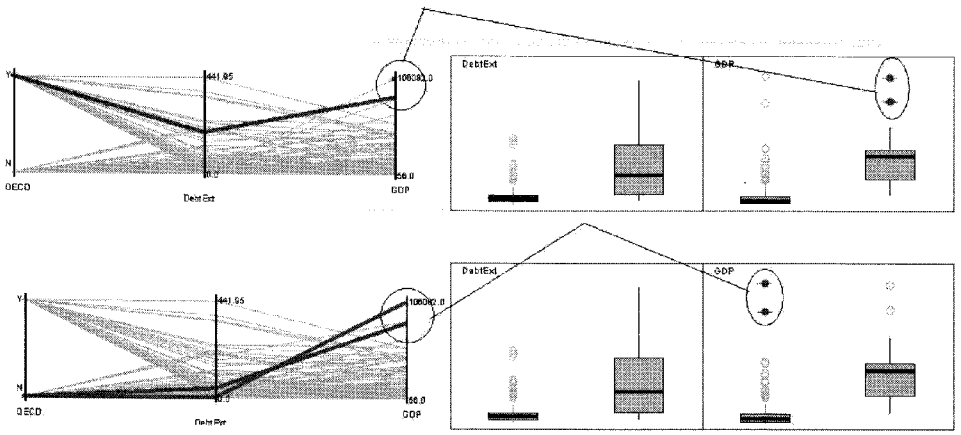

그림 3.6. OECD 가입국 중에서 GDP가 가장 높은 두 나라(LUXEMBURG 와 NORWAY)를, 아래에서는 비가입국 중 GDP가 가장 높은 두 나라 (LIECHTENSTEIN 과 QATAR)를 파란색으로 표시하였음. 가입국 중 LUXEMBURG는 외채가 결측값이고 비가입국 두 나라의 외채 대 GDP비율이 매우 낮은 것을 알 수 있음.

입국 중 외채가 1,843.7%로 가장 많은 나라(Monaco)와 다음으로 가입국이면서 외채가 매우 높은 나라(Ireland, 960.9%)가 있는데, 이들 두 나라 때문에 외채의 히스토그램의 모양이 매우 한쪽으로 치우쳐져 있어 분석하는 데 지장이 많은 것으로 나타난다. 이들 두 나라를 제거하고 다시 탐색을 시도해 본 것이 그림 3.6에 나타나 있다. 이 그림에서는 전반적으로 OECD 가입국은 비가입국에 비해 외채와 GDP가 높지만, 비가입국 중에서도 외채와 GDP가 가입국 못지않게 높은 나라들이 있다는 것을 알 수 있다. 또한 외채와 GDP 사이에는 별다른 연관성이 없는 것을 시각적으로 판단할 수 있다.

상자도형에서 OECD 가입국 중에서 특별하게 높은 GDP를 갖는 두 나라는 Luxemburg와 Norway로서 각각 104,452와 84,595이다. 이들 두 나라 중 Luxemburg는 외채가 결측값이고 Norway는 외채가 중간 정도이다. 우리나라의 경우, GDP는 19,836이고 외채는 18.25%이다. 또 비가입국 중 GDP가 가장 높은 두 나라는 Liechtenstein과 Qatar로서 GDP는 각각 10,6082, 83,152이며 외채는 각각 0%와 46.33%인 것으로 나타났다.

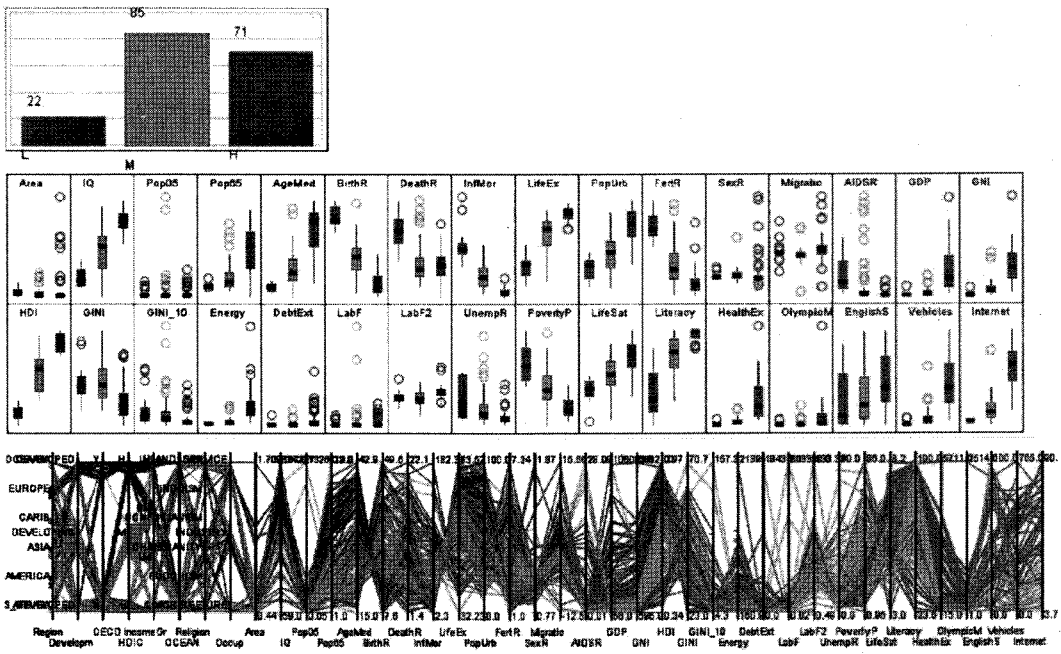

그림 **3.7.** HDIC와의 연관성을 파악하기 위한 평행좌표계와 평행 상자도형.

3.3. 인간개발지수와 가장 연관이 많은 특성들은 무엇인가?

이제 범주가 3개인 인간개발지수(HDIC)와 4개인 경제력 평가(incomeGroup)의 경우, 시각적 탐색 방법을 적용하여 보자. HDIC는 $(0,1)$ 사이의 값을 갖는 인간개발지수 HDI가 0.8 이상인 70개국을 HDIC = H, 0.502 이상인 85개국을 HDIC = M 그리고 나머지 22개국을 HDIC = L로 정의한 변수이다. 한편, HDI는 기대수명, 문자 해독률, 교육 정도 등을 기준으로 만들어 진것으로 한 국가의 웰빙척도로 이용되고 있다. 이는 UN 가입국 177개국을 상대로 측정하였으며, UN에서 지정하지 않은 3개 국가로 대만, Macao, Montenegro의 3개국이 있다. 이들 국가는 모두 HDIC = H로 분류되었다.

그림 3.7을 보면 HDIC와 다른 자원들과의 관계를 시각적으로 알아 볼 수 있다. 이 그림에서는 HDIC의 3개 범주를 색깔로 구분하고 (L-22개국 파란색; M-85개국 초록색, H-71개국 붉은색) 이를 평행좌표계와 평행 상자도형을 통해 탐색할 수 있다. 여기서 평행좌표계의 이산형 자원은 jittering을 하여 표현하였다. 이렇게 함으로써 여러 개의 관측값들을 따로 표현할 수 있고, 시각적 탐색이 가능해진다. 이를 보면 이산형 자원의 경우 HDIC는 지역, 종교, Ocean과는 큰 연관성이 없어 보인다. 그러나 incomeGroup은 HDIC와 매우 강한 연관성을 갖고 있고, developm와 OECD의 경우 HDIC = H인 경우와 그렇지 않은 경우가 구분되는 것으로 나타난다.

이제 연속형 자원을 탐색해보자. 평행좌표계의 경우, 각 국가의 자원들이 연결선분으로 나타나고 이들이 연결될 때 겹치는 것이 필연적이다. 따라서 선분이 많아지면 판독하기가 난해해 진다. 반면 상자도형은 각 범주별로 국가를 묶어 표현하기 때문에 선으로 나타나는 국가의 수가 아무리 많아도 표현하는 데 아무 지장이 없다. 따라서 상자도형이 평행좌표계보다 더 명확하게 구분하여준다. 이를 참고하여 HDIC와 관계가 많아 보이는 자원들을 살펴보면, IQ, AgeMed, BirthRate, InfMor, LifeEx, GNI, HDI, PovertyP, PopUrb, FertR, LifeSat, Literacy, HealthEx, vehicles, internet 등인 것으로 나타났

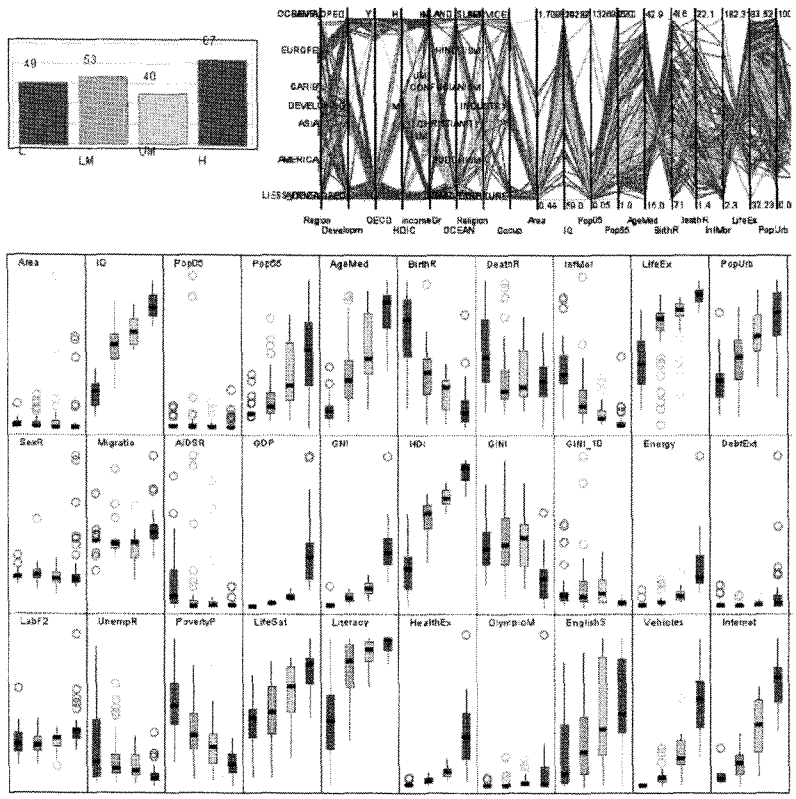

그림 **3.8.** 경제력지수와의 연관성을 파악하기 위한 평행좌표계와 평행 상자도형.

다. 여기서 DeathR, SexR, MigRatio, AIDSR, GINI, DebtExt, LabF, UnempR, EnglishS 등은 깊은 관계가 없는 것으로 나타난다. HDIC가 기대수명, 문자 해독률, 교육 정도 등을 기준으로 만들어진 것을 고려할 때 이와 관련된 자원이 강한 영향을 미치는 것은 당연한 결과라고 하겠다.

3.4. 경제력지수와 가장 연관이 많은 특성들은 무엇인가?

그림 3.8에는 4개의 범주를 갖는 경제력 지수(L: 49개국, LM: 53개국, UM: 40개국, H: 67개국)와 다른 자원과의 연관을 탐색할 수 있는 그림이 나타나있다. 평행좌표계는 이산형 자료를 포함한 처음 몇 개의 자원만 표시하였고, 상자도형에 연속형 자원들 모두가 incomeGroup의 각 범주별로 구분되어 나타나 있다. 평행좌표계를 보면 incomeGroup과 연관이 높은 이산형 자원은 역시 OECD 가입여부, HDIC, developm 등이고 Religion과 Occup은 연관이 별로 없는 것으로 나타났다.

한편, 연속형 자원을 살펴보면 IQ, Pop65, AgeMed, BirthR, InfMor, LifeEx, FertR, GDP, HDI, Energy, HealthEx, Vehicles, Internet 등과 IncomeGroup과의 연관이 많은 것으로 나타났으며, PopUrb, MigRatio, AIDSR, GINI, DebtExt, LabF, UnEmpR, PopvertyP, LifeSat, Literacy, OlynpicM, EnglishS 등과는 큰 관계가 없는 것으로 나타났다. 여기서 특이한 것은 생활만족도(LifeSat)와 문자 해독률(Literacy)이 경제력 평가지수와 크게 연관이 없다는 것이다. 이는 HDIC의 경우와 다른 점이라고 하겠다.

4. 연관성 척도에 의한 세계 각국 자원의 분석

도형을 사용하여 변수들간의 관계를 탐색하는 과정은 매우 직관적이고 자의적이다. 또한 시각적 방법을 통해 정보를 획득하는 과정은 숙달된 경험이 필요하다. 따라서 변수들간의 연관성에 대한 정보를 객관적으로 판단하기 위해서는 객관적인 연관성 척도가 필요하다.

변수들간의 연관성을 측정하는 기존의 통계학적 방법들은 변수들의 형식이 연속형 변수들이거나 이산형 변수들일 때에 유용하다. 예컨대, 가장 널리 이용되는 피어슨의 상관계수나 스피어맨의 순위 상관계수는 연속형 변수에 한해서 적용 가능하고, 이산형 변수와 이산형 변수의 연관성 척도로는 카이제곱 검정의 p-값을 활용할 수 있다. 또, 이산형 변수와 연속형 변수의 연관성을 측정하는 하나의 방법은 짝을 이루는 t-검정 등의 p-값을 사용할 수 있으며, 3개 이상의 카테고리를 갖는 경우, 분산분석 기법을 활용할 수 있다. 한편, 연속형과 연속형 변수의 연관성 측정에서 자료가 정규성을 따르지 않는다고 하면 Wilcoxon의 순위검정과 같은 비모수적 방법의 p-값을 사용할 수 있다. 그러므로 이들 p-값을 상호 비교함으로써 혼합형 자료의 연관성 척도에 대한 상대적인 크기를 비교할 수 있다 (Lee와 Huh, 2003). 그러나 여기서 문제가 되는 것은 각 검정에서 획득한 p-값이 일관성이 없다는 점이다. 예를 들어 X가 이산형이고 Y가 연속형인 경우, Wilcoxon의 검정에서 나타나는 p-값이 0.04이고 X와 Z가 모두 이산형일 때 카이제곱 검정에서 나타나는 p-값이 0.05였을 때 Y가 Z보다 더 X와 연관성이 높다고 할 수 있겠는가? 즉, 연속형과 이산형이 섞여있는 혼합변수들의 연관성을 일반적으로 판단하는 기준이 없었다. 그러나 Shannon의 엔트로피 이론에 기본을 둔 상호정보(mutual information: MI)는 매우 일반적인 연관성 측도기준으로서 혼합형 변수에 적용될 수 있는 장점이 있다.

4.1. MI의 추정

MI는 임의의 두 확률변수 X, Y의 결합분포함수와 주변분포함수가 각각 $P(x,y), P(x), P(y)$일때, 다음과 같이 정의된다.

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} dP(x,y) \log \frac{dP(x,y)}{d(P(x) \times P(y))}. \tag{4.1}$$

즉, MI, $I(X;Y)$는 X와 Y의 유형에 관계없이 정의될 수 있어, X와 Y가 어떠한 변수 집합이라고 하더라도 연관성을 일관성있게 측정할 수 있다.

MI는 엔트로피(entropy)에 의해 정의될 수 있다. 엔트로피는 확률변수의 불확실성을 의미하며, 연속형 확률변수의 엔트로피 $h(X)$는 다음과 같이 정의된다.

$$h(X) = \int_x dP(x) \log dP(x),$$

여기서 두 확률변수가 이산형인 경우, 적분은 합으로 표시된다. 즉,

$$h(X) = \sum_x p_x \log p_x \tag{4.2}$$

와 같다. 한편, MI와 엔트로피와의 관계는 다음과 같다.

$$\begin{aligned} I(X;Y) &= h(X) + h(Y) - h(X,Y) \\ &= h(X) - h(X|Y) = h(Y) - h(Y|X). \end{aligned}$$

MI는 이론적으로 연관성 측정을 위한 여러 장점이 있으나, 실제 응용에 있어서 문제가 되는 것은 MI의 추정이다. 이산형인 경우 p_x의 추정을 위해 우도추정량인 상대빈도를 사용하면 되지만, 연속형인 경우

확률밀도함수의 추정이 필요하다. 이 경우 표본 크기가 작으면 특히 많은 문제가 나타난다. 또한, 표본 크기가 크더라도 동점이 많이 나타나는 경우 추정에 어려움이 있다. 더욱이 수치적분할 경우, 적분의 범위를 어디까지 하고 몇 개의 점에서 값을 계산할 것인가에 대한 문제로 대두된다. 또, 커널추정을 사용하는 경우 어떤 커널을 사용하는 가에 따라 결과가 많이 차이가 날 수도 있다. 이러한 문제 때문에 연속형 자료를 이산화 시키고 이산화 된 결과를 사용하여 MI를 추정하기도 하지만, 이 또한 어떤 이산화 방법을 적용하는가에 따라 결과가 달라질 수 있다. 또, 이상값으로 인해 밀도함수가 한쪽으로 매우 심하게 치우쳐 있는 경우, 커널추정을 사용하여 적분을 구하면 매우 불안정한 추정값을 가져오게 된다.

4.2. DAVIS에 구현되어있는 MI 추정 방법

본 논문에서 다루는 문제들은 모두 목적변수가 이산형이고 설명변수는 이산형이나 연속형인 경우이다. 두 변수가 모두 이산형인 경우 (4.1)의 밀도함수는 각 셀 확률의 우도추정값인 상대빈도를 대입하여 MI를 추정하면 된다. 그러나 한 변수가 이산형이고, 다른 하나가 연속형인 경우 추정에서의 문제가 발생한다. DAVIS에는 이를 위한 다음과 같은 방법들이 구현되어있다.

1. 이산화 방법: 연속형 변수에 대해서는 이산화(discretization)을 먼저 하고 (4.2)를 적용시키는 방법으로, 여기서 이산화 하는 방법에는 동일 셀의 크기를 사용하는 방법, 히스토그램 방법, Fayyad와 Irani의 엔트로피 방법이 있다. 동일 셀의 크기는 전체 관측값의 수를 같은 상대도수를 갖는 여러 개의 셀로 나누는 것이다. 여기서 셀의 수는 사용자가 조정할 수 있다. 히스토그램 방법은 전체 범위를 몇 개의 동일한 너비로 나누는 것이다. 구간의 수는 히스토그램을 그리는데 추천되고 있는 Sturge의 공식을 사용한다. Fayyad와 Irani의 방법(이하 FI 방법)은 MDL(minimum description length) 원리에 의해 이산화를 시키는 방법으로서 구간의 수와 크기 등이 알고리즘에 의해 자동적으로 계산된다. 이 방법은 C4.5 결정나무를 작성할 때 연속형 변수를 이산화시키는 과정과 유사하다. 자세한 알고리즘은 Witten과 Frank (2005)에서 찾아볼 수 있다.
2. 적분법: 이 방법은 식 (4.1)에서 밀도함수를 추정하고, 다음과 같은 적분을 수치적분을 통해 구하는 방법이다.
$$\int_x \int_y d\widehat{P}(x,y) \log \frac{d\widehat{P}(x,y)}{d\left(\widehat{P}(x) \times \widehat{P}(y)\right)}.$$
3. 대입법: 이 방법은 식 (4.1)의 적분을 다음과 같이 근사한다.
$$\sum_x \sum_y d\widehat{P}(x,y) \log \frac{d\widehat{P}(x,y)}{d\left(\widehat{P}(x) \times \widehat{P}(y)\right)}.$$
2와 3, 두 방법에서 밀도함수의 추정은 kernel 방법을 사용하였고, kernel의 형식은 normal, Epanechnikov, uniform, biweight 등이 구현되어있다.
4. Sample-spacing 방법 (허문열과 차운옥, 2007).
5. k-nearest 방법 (차운옥과 허문열, 2008).

이 방법들은 모두 DAVIS에 구현되어 있으며 사용자가 임의로 방법을 선택할 수 있다. 저자들이 실험한 바에 의하면 여러 가지 방법 중 Fayyad와 Irani에 의한 이산화 방법을 적용하는 것이 자료의 종류에 큰 영향을 받지 않고 효율적인 것으로 나타났다. 따라서 앞으로 설명되는 모든 추정은 이 방법을 기준으로 한다.

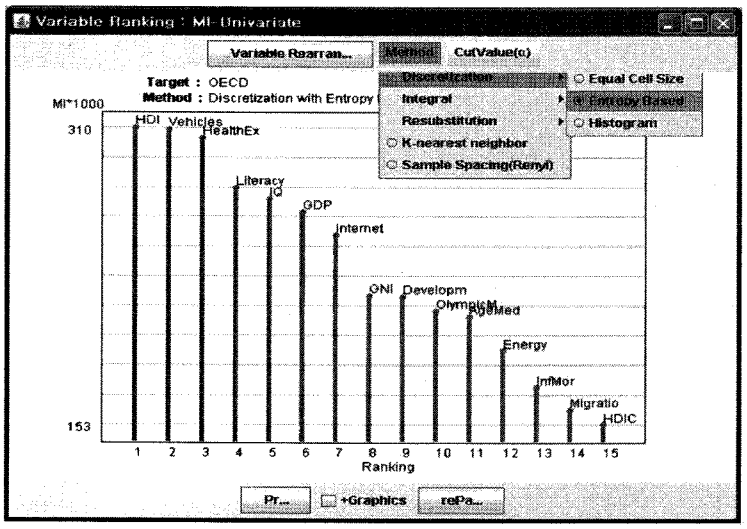

그림 **5.1.** MI 추정에 의해 OECD 가입여부와 가장 연관성이 높은 15개 자원.

5. MI에 의한 세계 각국의 자원 분석

이 절에서는 MI를 이용한 연관성 측정을 통해 다음과 같은 세 가지 문제에 대한 답을 구하려고 한다.

1. OECD 가입여부에 가장 연관이 많은 자원들은 어떤 것들 인가? 그리고 이들 변수 간에 어느 정도의 연관성이 있는가?
2. 인간개발 지수(HDIC)에 영향을 미치는 자원들은 무엇인가? 그리고 이들 변수 간에 어느 정도의 연관성이 있는가?
3. 국가 경제력(incomGr)에 영향을 미치는 자원들은 무엇인가? 그리고 이들 변수 간에 어느 정도의 연관성이 있는가?

5.1. OECD 가입여부와 가장 연관이 많은 자원들은 어떤 것인가?

그림 5.1에 OECD 가입여부와 각 자원의 MI를 추정한 결과가 나타나 있다. 이를 살펴보면 OECD 가입여부와 연관이 높은 자원으로는 HDI, 자동차 보급률, 보건지출비용이었으며 MI의 추정값이 각각 0.310, 0.309, 0.305로 거의 같아, 연관성에서 동일한 비중을 차지하고 있었다. 또, 문자해독과 IQ와 같이 지식과 관련된 항목도 높은 연관성이 있는 것으로 나타났으며, GDP, 인터넷 보급률, GNI 등의 경제력에 관한 항목도 다른 자원과 비교하여 연관성이 높았다.

우리나라와 일본은 어떤 위치에 있는가를 알아보기 위해 평행좌표계에 우리나라와 일본 두 나라를 하이라이트 시켜본 것이 그림 5.2에 나타나있다. 그림에서 붉은색은 한국, 파랑색은 일본을 나타낸다. 이를 보면 우리나라는 문자해독율과 IQ는 전 세계에서 1, 2위에 속하지만 그 외의 자원들은 그다지 높지 않은 것을 알 수 있다. 즉, HDI가 전체 국가 중에서 상위권이기는 하지만, OECD 가입국 중에서는 그다지 높지 않고, 자동차 보급률도 중간 정도이다. 또, 인터넷 보급률도 정보강국으로 알려진 것에 비하면 그다지 높지 않았다. 그리고 GDP, GNI, energy 소비량 등의 경제력도 아직 미흡한 상태이다. 한편, 일본은 인간개발지수가 매우 높고, 자동차 보유대수, GDP, GNI 등이 높다. 이 그림에 의하면 이상값이

표 5.1. 인구연령 중앙값 비교

	한국	일본	중국	미국	개발국	세계
Population 05	48	128	1,313	300	1,216	6,514
Population 50	42	103	1,409	402	1,245	9,191
ageMed 05	35.0	42.9	32.5	36.0	38.6	28.5
ageMed 50	54.9	54.9	45.0	41.1	45.7	38.1
%age14– 05	18.6	13.9	21.6	20.8	17.0	28.3
%age14– 50	10.4	11.3	15.3	17.3	15.2	19.8
%age65+ 05	9.4	19.7	7.7	12.3	15.3	7.3
%age65+ 50	35.1	37.7	23.7	21.0	26.1	16.2

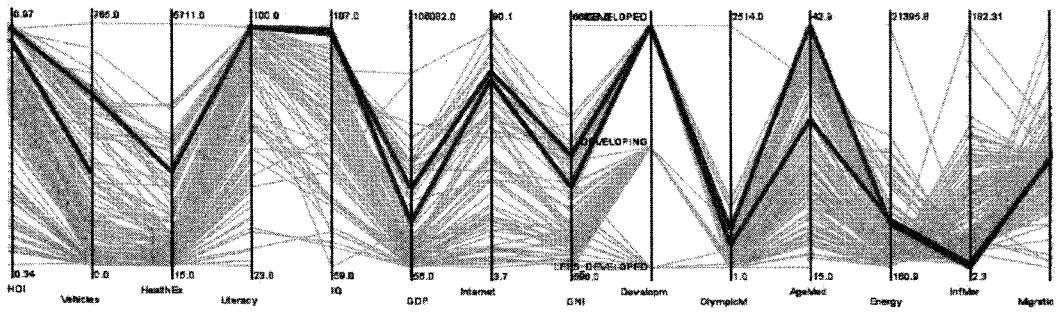

그림 5.2. OECD 가입여부와 연관성이 높은 자원의 한국과 일본 비교.

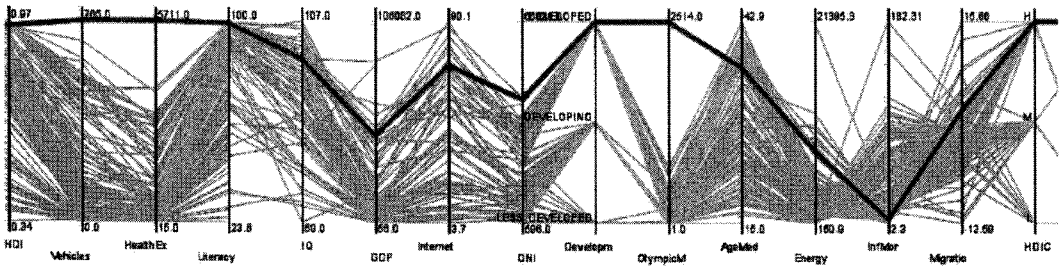

그림 5.3. 미국의 위치

몇 개 있어 일본이 그다지 높지 않은 것처럼 나타나지만, 이들 몇 개를 제외하면 일본은 높은 그룹에 속한다. 참고로 그림 5.3을 보면 미국이 어느 위치에 있는가를 알아 볼 수 있다. 미국은 IQ, Internet 보급률, 인구연령 중앙값 등은 아주 상위권에 속하지 않으나, 그 외에는 전부 상위권에 있다는 것을 알 수 있다.

2005년도 인구센서스에 의하면 그림에서와 같이 한국의 인구연령 중앙값은 35세로 일본의 43세보다 낮았다. 그러나 2050년이 되면 한국과 일본 모두 인구연령의 중앙값이 55세가 된다고 한다. 표 5.1에 한국, 일본, 중국, 미국, 개발국 그리고 전 세계에 대해 UN에서 조사 및 예측한 연령중앙값, 14세 이하 인구비율, 65세 이상 인구비율 등이 나타나있다.

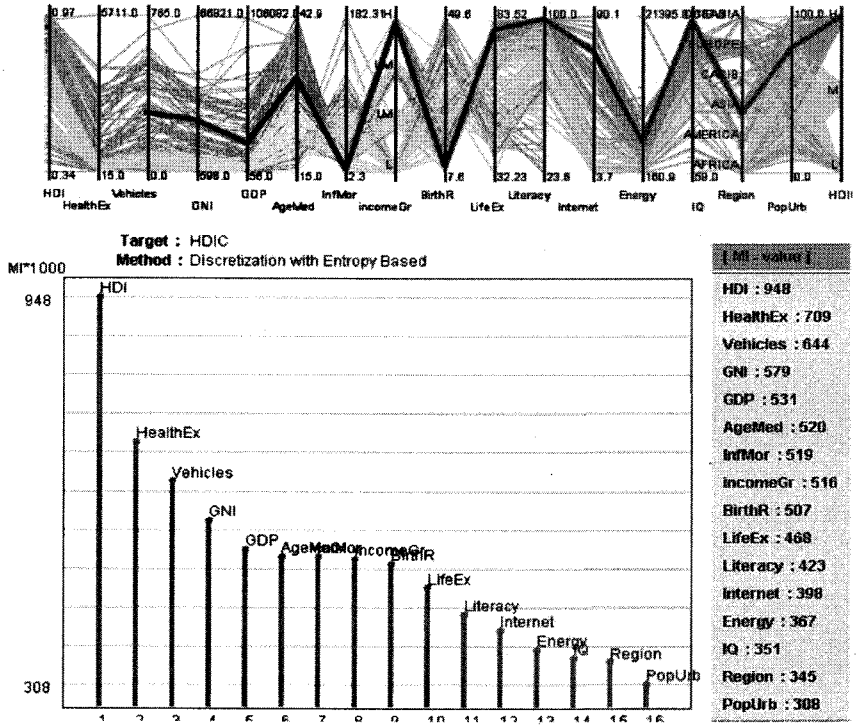

그림 5.4. MI 추정에 의해 HDIC와 가장 연관성이 높은 15개 자원

5.2. 인간개발 지수, 국가 경제력, 개발정도 등과 가장 연관이 많은 자원들은 어떤 것인가?

OECD 가입여부는 범주가 두 개인 경우인데 반해, 인간개발지수(HDIC)와 개발 정도(Developm)는 3개의 범주를 갖는다. 또, 국가 경쟁력은 4개의 범주를 갖고 있다. 이 세 자원에 가장 연관이 많은 변수들을 비슷한 방법에 의해 분석한 결과가 그림 5.4, 5.5, 5.6에 나타나 있다. 우리나라의 경우를 세계적인 수준과 비교하기 위해 그림에서는 붉은색으로 표시하였다.

HDIC는 기대수명, 문자해독율, 교육 정도, 그리고 GDP 등을 종합적으로 고려하여 작성한 한 국가의 웰빙지수인 HDI를 상(0.801 이상 73개국), 중(0.502 이상 85개국), 하(0.502 미만 22개국) 3개 범주로 구분한 것이기 때문에 본 논문에서 사용하고 있는 FI 방법에 의한 MI 추정이 잘되었다면 이들과 관련된 항목들이 높은 값으로 추정되어야 할 것이다. 실제 결과를 보면 HDI, 건강 비용, 자동차 보급률, GNI와 GDP, 다음으로 국민 연령, 유아 사망률, 기대수명 등과 같은 수명에 관한 자원과 문자 해독률, 인터넷 보급률 등 교육에 관련된 내용이 높은 연관을 갖고 있다. 따라서 이 추정 방법이 효율적이라는 것을 실증적으로 보여주고 있다.

마지막으로 한 국가의 개발 정도(Developm)에 어떤 자원들이 가장 많이 연관되어 있는가를 살펴보기로 한다. 연관이 높은 순서대로 나열하면, 인터넷 보급률(0.54), 유아 사망률(0.51), IQ(0.50), 인간개발지수(0.49), 보건지출비용(0.44), 기대수명(0.44), 인구연령 중앙값(0.36), 지역(0.33), 문자 해독률(0.32), 경제력(0.32), 65세 이상 인구비율(0.29), 여성 가임률(0.26), 노동력(0.14), 외채(0.13), 빈곤층 비율(0.12)이었다. 즉, 국민의 건강과 지적 능력에 관한 부분이 높은 영향을 미치는 것으로 나타났다.

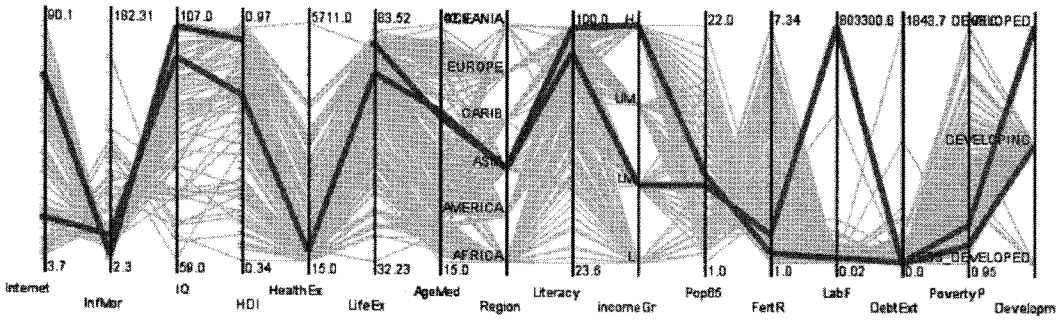

그림 **5.5.** MI 추정에 의해 HDIC와 가장 연관성이 높은 15개 자원

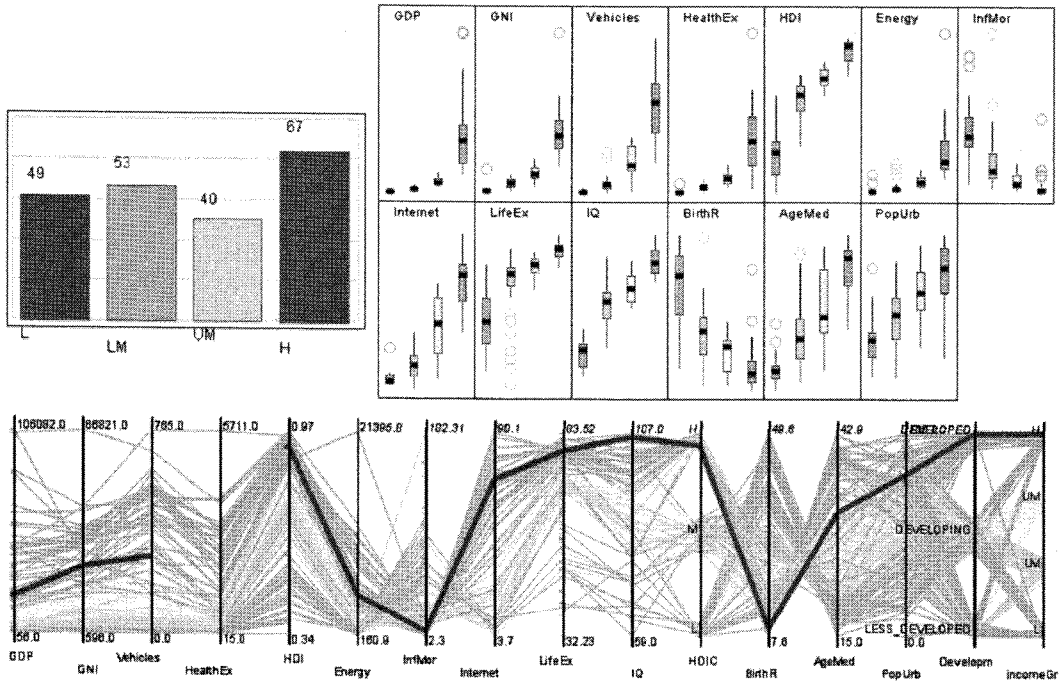

그림 **5.6.** MI 추정에 의해 HDIC와 가장 연관성이 높은 15개 자원

우리나라를 개발국 또는 개발도상국으로 분류하느냐는 기준을 작성하는 기관에 따라 차이가 있다. 실제로 개발국을 정의할 때 북아메리카와 유럽의 국가 및 아시아에서는 일본이라고 정의하는 기관도 있다. 본 논문에서 사용하는 자료는 우리나라를 개발국으로 구분하였다. 그림 5.5에서는 개발 정도에 영향을 많이 미치는 자원들에 대해 우리나라가 세계에서 어떤 수준인가를 보여주고 있다. 이를 보면 유아생존률, IQ, 문자 해독률 등에서 세계의 정상급에 속하고, 인터넷 보급률과 HDI는 아직 최상급이 아니지만 높은 편에 속한다. 또한 수명과 65세 이상 인구비율은 이 자료가 주어진 시점 (UN 자료, 2005년 기준)에서는 높지 않지만 앞에서 살펴본 바와 같이 2050년에는 세계에서 가장 높은 나라가 된다. 따라서 우리나라는 향후 매우 수준이 높은 나라가 될 것으로 기대된다.

마지막으로 국가 경제력에 대한 내용을 살펴 본 것이 그림 5.6에 나타나 있다. 이를 보면 처음 7개가 GDP, GNI, 자동차 보유대수, 보건지출비용, HDI, 에너지 소비, 인터넷 보급률 등 경제와 관련된 항목이고 다음 항목들이 기대수명, IQ, 출생률, 국민연령 등 인구관련이다. 우리나라의 경우, 인구관련 내용은 매우 상위에 속하지만 경제관련 내용은 아직 높지 않은 것으로 나타났다.

6. 결론

본 논문에서는 세계 232개국의 40여개 자원을 사용하여 OECD 가입 여부, 국가의 웰빙, 개발정도 및 경제력에 가장 영향을 많이 미치는 것들이 무엇인가를 살펴보았다. 즉, 한 나라가 잘 살기 위해서는 어떤 자원들이 가장 중요한 요인인가를 살펴 보았다. 여기서 획득한 자료는 이산형과 연속형이 복합적으로 섞여 있으며, 결측값 및 이상값이 많아 기존의 통계적 방법을 적용하는 것이 용이하지 않다. 본 논문에서는 데이터 시각화 방법을 적용하여 자료에 내재되어 있는 정보를 탐색하는 접근법을 제안하였다. 그러나 데이터 시각화 방법도 자료의 형태가 복잡하고 이상값이 많아지면 탐색하는 방법도 매우 숙달된 훈련이 필요할 뿐만 아니라 결과를 해석하는 과정도 용이하지 않은 것을 알 수 있었다.

본 논문에서는 혼합형 변수들간의 연관성을 추정하는 상호정보 개념을 도입하여 이 문제를 해결하였다. 이를 사용함으로서 한 국가의 개발정도나 경제력, 웰빙 그리고 OECD 가입 여부와 각 자원과의 연관성을 측정할 수 있었다. 이렇게 하여 추정된 연관성의 크기에 따라 자원들을 순서 배열하고 이 중에서 가장 높은 몇 개의 자원만 택하여 다시 시각적 탐색 방법을 적용함으로서 자료분석 결과가 객관적이면서 이 해석도 용이해졌다. 이를 수행해 본 결과 IQ, 출생률, 유아 사망률, 국민연령, 문자 해독률 등 국가의 인구관련 자원과 자동차 보급률, 인터넷 보급률, GDP, GNI 등 경제관련 자원이 중요한 것을 알 수 있었다. 우리나라의 경우 인구관련 내용은 아주 상위권에 속하지만 경제관련 내용은 아직 많이 발전시켜야 할 여지가 있는 것으로 나타났다. 또한 중국을 보면 하나 낳기 운동을 벌여 유아 사망률 등 인구관련 내용이 매우 상위권에 속한다. 더욱이 13억에 이르는 인구의 IQ가 매우 높은 국가(중국 100, 한국 106)로서 비록 현재 상태로서는 경제관련 자원들이 낮아 아직 개발 정도와 웰빙 등의 기준으로 볼 때 많이 부족하지만 향후 매우 발전가능한 나라인 것으로 나타났다.

참고문헌

차운옥, 허문열 (2008). 상호정보 추정을 위한 k-최근접이웃 기반방법, <한국통계학회논문집>, **15**, 977–991.

허문열, 차운옥 (2007). 정보이론과 시각화 방법에 의한 여론조사 분석의 새로운 접근방법, <응용통계연구>, **20**, 61–78.

Fayyad, U. M. and Irani, K. B. (1992). On the handling of continuous-valued attributes in decision tree generation, *Machine Learning*, **8**, 87–102.

Huh, M. Y. and Song, K. R. (2002). DAVIS: A Java-based data visualization system, *Computational Statistics*, **17**, 411–423.

Lee, S.-C. and Huh, M. Y. (2003). A measure of association for complex data, *Computational Statistics & Data Analysis*, **44**, 211–222.

Shannon, C. E. A. (1948). Mathematical theory of communication, *The Bell Systems Technical Journal*, **27**, 379–423, 623–656.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition, Morgan Kaufmann, San Francisco.

Statistical Consideration on the Resources of the Countries in the World

Moon Yul Huh[1] · Byong Su Choi[2] · Seung-Chun Lee[3]

[1]Dept. of Statistics, Sungkyunkwan University; [2]Dept. of Mutimedia Engineering, Hansung University; [3]Dept. of Statistics, Hanshin University

(Received November 2008; accepted December 2008)

Abstract

The paper investigates the resources of the 232 countries based on the 39 resources of these countries. The data used in this work is from various sources like UN, CIA, World bank, OECD reports and the home pages of each country. The purpose of the study is to evaluate what resources are most influential to the wealth of a country, to the well-being of the country, or the status of the country's development. For this, data visualization method is applied. Data visualization technique, although powerful for exploratory purposes, is dependent upon the users expertize and the interpretation is also dependent on the of the users. For objective methods of investigation, mutual information based on the Shanon's entropy theory is applied here. All the statistical methods employed in this paper are processed with DAVIS (Huh and Song, 2002).

Keywords: Parallel coordinates, parallel box plot, data visualization, mutual information, DAVIS.

This Research was financially supported by Hansung University in the year of 2007.

[3]Corresponding author: Professor, Dept. of Statistics, Hanshin University, 411 Yangsan-Dong, Osan, Kyunggi-Do 447-791, Korea. E-mail: seung@hs.ac.kr