전이행렬자료의 동적 단순대응분석

서명록 [1] 최용석[2] 강창완[3] 임승범[4]

요 약

일반적으로 단순대응분석에서는 하나의 분할표 자료에 대한 행과 열의 대응관계만을 주로 다루어 왔으나 시점의 변화에 따른 행과 열 범주의 대응관계에 대한 변화의 추세를 나타내지는 못했다. 본 연구에서는 새로이 추가범주를 활용한 전이행렬자료의 동적 단순대응분석(dynamic simple correspondence analysis of transition matrix data: DSCA)을 제안하고자 한다.

DSCA는 시점의 변화에 따른 행과 열 범주의 변화되는 대응관계뿐만 아니라 행 범주들의 시간적인 변화의 경향을 보여주는 장점을 갖고 있다. 또한 기준시점에서 다음 시점으로의 변화도 예측하여 보여줌으로써 향후 변화의 경향을 시각적으로 보여준다.

주요용어 : 동적 단순대응분석, 마르코프 모형, 전이행렬, 추가범주

1. 서론

고정된 패널을 대상으로 여러 번 측정된 동일한 개수의 범주에 대하여 시간의 흐름에 따른 각 패널의 범주 사이 전이 모습을 전 시점과 현 시점의 2차원 분할표로 정리한다. 이러한 분할표를 전이행렬자료(transition matrix data)라 하며, 전이행렬자료를 통하여 동일한 범주에 머무르는 유지정도와 다른 범주로의 전이정도를 살펴볼 수 있다. 또한 여러 개의 전이행렬자료가 있는 경우 마르코프 모형(one-step markov model)을 통하여 전이확률을 구할 수 있고, 구하여진 전이확률을 통하여 다음 시점의 전이행렬자료도 예측이 가능하다(Bishop 외 2인, 1975, pp.257~267). 또한, 범주형 자료의 분석을 위한 통계적 기법으로 비정칙치 분해(singular value decomposition)를 이용한 차원축소와 함께 2차원의 그래프적 표현을 통해 분할표 자료의 행과 열 범주들간의 대응관계를 탐구하려는 대응분석이 있다(Greenacre, 1994, pp. 3~8; 최용석, 2001, 1장). 대응분석은 이원분할표에 대한 단순대응분석과 다원분할표에 대한 다중대응분석이 있다. 지금까지의 단순대응분석에서는 하나의 분할표 자료에 대한 행과 열의 대응관계만을 주로 다루어 왔으나 시점의 변화에 따른 행과 열 범주의 대응관계에 대한 변화의 추세를 나타내지는 못했다. 본 연구에서는 새로이 추가범주를 활용한 전이행렬자료의 동적 단순대응분석(dynamic simple correspondence analysis of transition matrix data: DSCA)을 제안하고자 한다. DSCA는 전이행렬자료에 대해서 단순대응분석을 실시하되 기준이 되는 분할표 자료행렬에 시점의 변화에 따라 추가되는 전이행렬자료의 행과 열의 범주를 추가범주(supplementary category)로 지정한다. 이것은 시점의 변화에 따른 행과 열 범주의 변화되는 대응관계뿐만 아니라 행 범주들의 시간적인 변화의 경향을 보여주는 장점을 갖고 있다. 또한 기준시점에서 다음 시점으로의 변화도 예측하여 보여줌으로써 향후 변화의 경향을 시각적으로 보여준다.

1) 부산대학교 자연과학대학 통계학과 석사과정
2) 부산대학교 자연과학대학 통계학과 교수
3) 동의대학교 자연과학대학 정보통계학과 교수
4) 동의대학교 자연과학대학 정보통계학과 석사과정

2. 추가범주를 활용한 전이행렬자료의 동적 단순대응분석

어떤 분할표의 대응분석에서 범주들의 대응관계를 제공한 2차원의 대응분석 그림에다 새로운 정보를 가진 범주를 추가하여 나타내는 기법은 매우 중요하다. 이를 추가범주를 활용한 대응분석이라 한다. 이 때 기존의 범주를 활동적 요소(active element)라 하고 추가되는 범주를 설명적 요소(illustrative element)라 부른다(Greenacre, 1984, pp. 70~76; Lebart et al. 1984, p. 42). 크기가 $n\times p$인 자료행렬 O에 n_s개의 행과 p_s개의 열이 추가되었다고 하면 <그림 2.1>과 같이 나타난다. 여기서 O_+는 크기가 $n_s\times p$인 추가행자료행렬이고 O^+는 크기가 $n\times p_s$인 추가열자료행렬이다. 일반적으로 추가범주에 대한 저차원 공간상의 좌표점을 구하는 알고리즘은 아주 다양하다. 여기에서는 Greenacre(1984, pp. 71~73)의 알고리즘을 소개하려 한다.

n개의 행프로파일 $r_1,\cdots,r_n$에 의한 행렬 R과 p개의 열프로파일 $c_1,\cdots,c_p$에 의한 행렬 C는 다음과 같이 구성될 수 있다.

$$R=D_r^{-1}F=(r_1,\cdots,r_n)',\quad C=D_c^{-1}F'=(c_1,\cdots,c_p) \tag{2.1}$$

식 (2.1)처럼 추가행프로파일행렬 R_s와 추가열프로파일행렬 C_s도 다음과 같이 구해진다.

$$R_s=diag(O_+1_p)^{-1}O_+',\quad C_s=diag(O^{+\prime}1_n)^{-1}O^{+\prime} \tag{2.2}$$

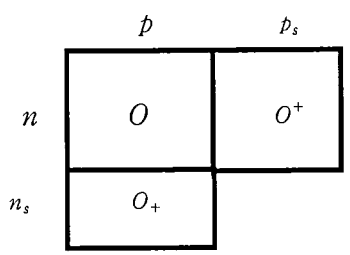
식 (2.2)에서 $diag(\ vector)^{-1}$는 벡터를 대각행렬로 만들고 그것의 역행렬을 구함을 말한다.

추가범주가 없는 단순대응분석을 위한 행과 열좌표점을 각각 $X=D_r^{-1}AD_u$ 와 $Y=D_c^{-1}BD_u$라 하자. 여기서 A와 B는 대응행렬 F의 일반화비정칙값분해를 만족하고 다음의 조건식 $A'D_r^{-1}A=B'D_C^{-1}B=I$ 을 따른다. 이 조건과 식 (2.1)에 의해서 행과 열좌표점은 각각 다음과 같다.

$$\begin{aligned}X&=D_r^{-1}AD_u(B'D_c^{-1}B)=(D_r^{-1}F)D_c^{-1}B=RD_c^{-1}B\\ Y&=D_c^{-1}BD_u(A'D_r^{-1}A)=(D_c^{-1}F')D_r^{-1}A=C'D_r^{-1}A\end{aligned} \tag{2.3}$$

추가행과 열프로파일행렬의 저차원 공간상의 좌표점을 각각 X_s와 Y_s라 하면 다음과 같다.

$$X_s=R_sD_c^{-1}B\ ,\quad Y_s=C_sD_r^{-1}A \tag{2.4}$$



<그림 2.1> 추가된 행과 열 범주의 표시

여기서 조사의 시작 시점의 전이행렬자료를 O라 하고 시간의 흐름에 따라 추가되는 전이행렬 자료들을 O_+의 추가행자료나 O^+의 추가열자료로 처리하여 2차원상에 투영시켜서 시점의 변화에 따른 행과 열 범주들간의 동적인 변화 모습을 살펴볼 수 있다. 이러한 기법을 본 연구에서는 추가범주를 활용한 전이행렬자료의 동적 단순대응분석(DSCA)라 하겠다.

3. 응용사례

매월 발표되는 통계청의 고용동향 분석에서 실업은 경제지표로서 중요한 역할을 하고 있다. 그러나 실업의 구성비율이나 발생경로, 재취업 경로와 비경제 인구의 흐름에 대한 고용동향 정보는 부족하다. 본 연구에서는 통계청에서 제공하는 2002년 경제활동인구를 대상으로 전이행렬 자료와 추가범주를 활용한 동적 단순대응분석으로 다각적, 시각적인 분석을 하고자 한다.

3.1 월간 고용동향분석

<표 3.1.1> 2002년 1월과 2월의 전이행렬자료

1월 \ 2월	A	B	C	D	E	F	X	Z	합
A	3383	10	13	11	6	2	3	122	3550
B	4	5583	4	17	12	4	30	105	5759
C	7	2	1863	6	6	3	25	90	2002
D	9	16	2	7729	28	15	42	186	8029
E	3	12	7	25	7156	5	29	119	7356
F	2	1	4	14	3	2839	10	35	2908
X	0	27	41	62	43	17	614	110	914
Z	145	100	93	203	185	32	164	21527	22449
합	3553	5751	2027	8067	7439	2917	917	22396	52967

<표 3.1.1>의 전이행렬자료에서 대각행렬은 각 범주별로 유지하는 개체의 수를 의미하고, 행은 각 범주별로 다른 범주로 전출한 개체의 수를 열은 다른 범주로부터의 전입한 개체의 수를 의미하며 각 영문자는 고용구분을 나타내는데 다음 <표 3.1.2>와 같다.

<표 3.1.2> 고용구분표

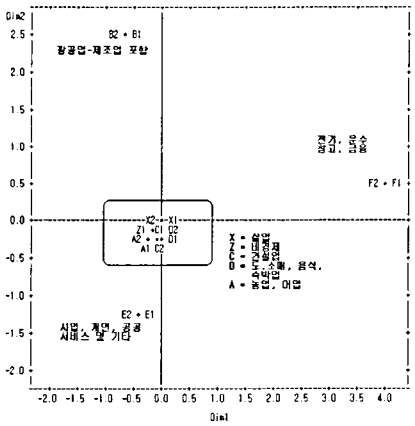
표기	고용구분
A	농업, 어업
B	광공업(제조업포함)
C	건설업
D	도.소매, 음식, 숙박업
E	사업, 개인, 공공서비스 및 기타
F	전기, 운수, 창고, 금융
X	실업
Z	비경제 활동인구

<표 3.1.1>에서 첫째, 각 범주별 유지정도는 대각행렬을 보면 알 수 있는데 대체로 90%이상을 유지하고 있으며, X(실업)의 유지 정도는 67%인 것으로 나타났다. 둘째, 전이행렬자료의 행들은 전출을 의미하는 것으로 A~X까지의 경제활동인구의 변동에 있어서 Z로의 전출이 가장 많이 나타나고 있다. 또한 Z의 열은 전입을 의미하는 것으로 전출과 같은 경향을 보이고 있다. 셋째, X(실업)를 기준으로 전입은 Z에서 가장 많이 오며, 전출 또한 그러하다. 이것은 실업자가 비경제활동인구가 되고 다시 비경제활동인구가 실업자가 되는 것으로 취업자가 되지 못하는 악순환을 보이고 있음을 시사한다. 또한 A(농업,어업)에서 실업자가 되는 경우는 거의 드물며, 실업자가 A로 취업을 하는 경우는 없는 것으로 나타났다. 끝으로, X와 Z를 제외한 A에서 F까지의 각 범주별 전출(행)과 전입(열)을 살펴보면 산업별 이직현상을 관찰할 수 있다. A는 주로 C로 전출하며 C, D에서 전입을 한다. B는 D, E로 전출을 하며 D, E에서 전입을 많이 하

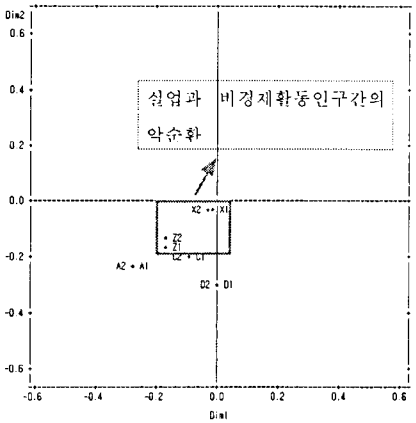
는 것을 볼 수 있다. 기타 나머지 산업도 이러한 해석이 가능하다.

<그림 3.1.1>과 <그림 3.1.2>는 <표 3.1.1>을 단순대응분석 한 것이다. <그림 3.1.1>에서 제1축(Dim1)을 기준으로 보면 1사분면의 F(전기, 운수, 창고, 금융)는 X(실업), Z(비경제)와 대응관계를 보이지 않고 있음을 알 수 있다. 여기서 F는 실업과 비경제의 발생이 다른 산업에 비해 적다는 것을 알 수 있다. 2사분면의 B(광공업-제조업 포함)는 X(실업), Z(비경제)와 대응관계를 보이지만 거리가 상당히 멀어서 다른 산업에 비해 실업과 비경제가 덜 발생함을 알 수 있다.

<그림 3.1.2>는 <그림 3.1.1>의 □(사각형) 부분을 더 세분화해서 본 것이다. 제1축(Dim1)을 기준으로 보게되면 모든 좌표 점들이 왼쪽에 위치하고 있다. 특히 X(실업)와 Z(비경제)가 가까이 대응하고 있으므로 둘 사이의 밀접한 관련성을 알 수 있고, 이것은 앞에서 언급된 실업과 비경제의 악순환이 나타나고 있음을 보여준다. 또한 X(실업)에서 C(건설업), D(도.소매, 음식, 숙박업), E(사업, 개인, 공공서비스 및 기타)로 재취업을 많이 하는 것을 알 수 있다. 그리고 Z(비경제)와 C(건설업)와 A(농업, 어업)가 대응관계가 강한 것을 볼 수 있다. 결론적으로 X에서 C, D, E로 재취업을 많이 하고 Z에서는 C, A로 취업을 많이 함을 알 수 있다.



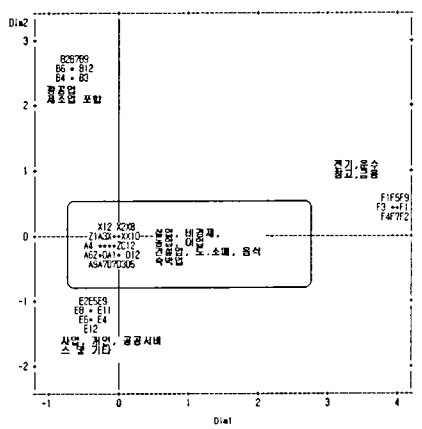
<그림 3.1.1> <표 3.1.1>의 단순대응분석그림



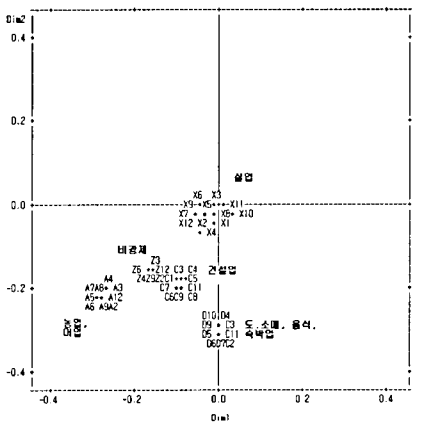
<그림 3.1.2> <그림 3.1.1>의 □ 부분 세분화한 그림

3.2 연간 고용동향분석

<그림 3.2.1>에서 제1축(Dim1)을 기준으로 보면 1사분면의 F(전기, 운수, 창고, 금융)는 10월과 11월에 실업과 대응관계를 보이고 있다. 반면 2사분면의 B(광공업-제조업 포함)는 10월과 11월을 제외한 나머지 10개월 동안 실업과 대응관계를 보이고 있다. <그림 3.2.2>를 보면 그림의 중앙에 위치한 X(실업)는 10월과 11월을 제외한 나머지 10개월 동안에 3사분면의 Z(비경제)와 가장 가깝게 대응하고 있다. 이것은 10개월 동안의 실업과 비경제 사이의 악순환을 보여주는 것이다. X중심으로 취업하는 형태를 보면 C(건설업)나 D(도.소매, 음식, 숙박업)로 많이 가는 것을 볼 수 있다. 또한 Z중심으로 보면 1순위는 C, 2순위는 A(농업, 어업)로 가는 것을 볼 수 있다. 또한 실업의 발생을 보면 C, D, E(사업, 개인, 공공서비스 및 기타), B에서 F(전기, 운수, 창고, 금융)에 비해 많이 발생함을 알 수 있다. 위의 내용을 정리하면 2002년 전체 고용동향에서의 문제점은 첫째, 실업인구가 건설이나 도소매, 음식, 숙박업, 사업, 개인, 공공서비스 및 기타로 취업을 하게 되지만 또다시 실업이 되는 악순환의 연속을 보인다는 것과 둘째, 오랜 실업으로 재취업을 포기하고 비경제인구로 전락한다는 것이다.



<그림 3.2.1> 2002년 전체 고용동향 그림



<그림 3.2.2> <그림 3.2.1>의 □ 부분 세분화한 그림

3.3 마르코프 모형을 이용한 고용동향 예측

<표 3.3.1> 1월에서 11월까지의 누적도수 및 전이확률

	A	B	C	D	E	F	X	Z	합
A	44833	129	149	128	91	34	26	1226	46616
	(96.2)	(0.3)	(0.3)	(0.3)	(0.2)	(0.1)	(0.1)	(2.6)	(100)
B	140	55447	71	189	144	41	301	1165	57498
	(0.2)	(96.4)	(0.1)	(0.3)	(0.3)	(0.1)	(0.5)	(2.0)	(100)
C	144	69	20861	81	75	33	224	756	22243
	(0.6)	(0.3)	(93.8)	(0.4)	(0.3)	(0.1)	(1.0)	(3.4)	(100)
D	129	205	100	76507	324	125	476	2301	80167
	(0.2)	(0.3)	(0.1)	(95.4)	(0.4)	(0.2)	(0.6)	(2.9)	(100)
E	109	125	74	270	73338	85	367	1562	75930
	(0.1)	(0.2)	(0.1)	(0.4)	(96.6)	(0.1)	(0.5)	(2.1)	(100)
F	32	27	31	103	79	28499	116	342	29229
	(0.1)	(0.1)	(0.1)	(0.4)	(0.3)	(97.5)	(0.4)	(1.2)	(100)
X	27	320	350	607	515	146	4738	842	7545
	(0.4)	(4.2)	(4.6)	(8.0)	(6.8)	(1.9)	(62.8)	(11.2)	(100)
Z	2393	1186	993	2252	1742	298	1041	200637	210442
	(1.1)	(0.6)	(0.5)	(1.1)	(0.8)	(0.1)	(0.5)	(95.3)	(100)
합	47807	57508	22629	80137	76306	29261	7289	208731	529670

<표 3.3.2> 예측된 11월과 12월의 전이행렬자료

11월 \ 12월	A	B	C	D	E	F	X	Z	합
A	4560	13	15	13	9	3	3	125	4741
B	14	5563	7	19	14	4	30	117	5769
C	15	7	2240	9	8	4	24	81	2388
D	13	20	10	7634	32	12	47	230	7999
E	11	13	8	28	7470	9	37	159	7734
F	3	3	3	10	8	2867	12	34	2940
X	2	28	31	53	45	13	413	73	658
Z	236	117	98	222	172	29	103	19762	20738
합	4855	5764	2411	7987	7759	2941	669	20581	52967

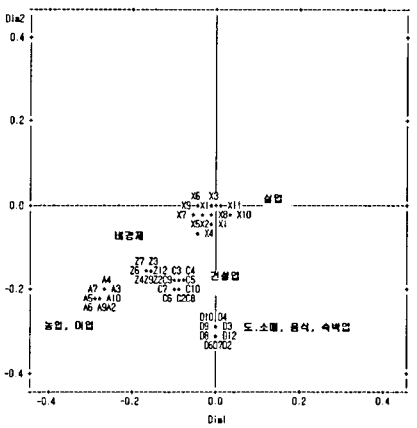
<표 3.3.3> 실제 11월과 12월의 전이행렬자료

11월 \ 12월	A	B	C	D	E	F	X	Z	합
A	3892	35	14	8	8	0	8	776	4741
B	7	5596	2	8	4	5	25	122	5769
C	7	0	2266	1	2	0	17	95	2388
D	8	7	2	7769	4	5	44	160	7999
E	6	3	4	18	7567	1	29	106	7734
F	2	1	1	3	6	2902	3	22	2940
X	1	24	15	39	26	15	501	37	658
Z	48	60	49	183	101	20	83	20194	20738
합	3971	5726	2353	8029	7718	2948	710	21512	52967

<표 3.3.1>은 통계청에서 제공한 2002년 경제활동인구 자료의 1월에서 11월까지 각 산업별 전이행렬자료의 누적도수 및 마르코프 모형을 통해서 계산한 전이확률표이고 <표 3.3.2>는 <표 3.3.1>의 전이확률값을 이용하여 만든 11월과 12월의 전이행렬자료이다.

<표 3.3.3>은 2002년 11월과 12월의 산업별 변화의 실제 전이행렬자료이다. <표 3.3.2>와 <표 3.3.3>을 비교하면 A(농업,어업)와 Z(비경제활동인구)를 제외하고는 전반적으로 같은 경향이 나타났다. <그림 3.3.1>과 1월부터 12월의 실제자료를 이용하여 동적 단순대응분석을 하였던 <그림 3.2.2>를 비교하면 각 산업의 변화 추이를 나타내는 좌표점이 정확하게 같은 위치에 자리하지는 않지만 그룹을 이루는 위치 내에 들어가 있음을 알 수 있고 여기서 고용동향의 흐름을

마르코프 모형을 이용하여 예측 가능함을 알 수 있다.



<그림 3.3.1> 예측된 전이행렬자료를 이용한 동적 단순대응분석그림

4. 결론

본 연구에서는 기존의 전이행렬자료의 분석에서 다루지 못했던 부분을 추가범주를 활용한 동적 단순대응분석을 활용하여 전이행렬자료의 변화모습을 다각적, 시각적으로 살펴보았다. 특히 향후 전이행렬자료의 변화 패턴도 예측할 수 있었다.

참고문헌

[1] 최용석 (2001). <SAS 대응분석의 이해와 응용>, 자유아카데미, 서울.

[2] Greenacre, M. and Blasius, J. (1994). *Correspondence Analysis in the Social Sciences*, Academic Press, London.

[3] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis : Theory and Practice*, The MIT Press, London.

[4] Lebart, L., Morineau, A. and Warwick, K.(1984). *Multivariate Descriptive Statistical Analysis : Correspondence Analysis and Related Techniques for Large Matrices*, Wiley, New York.