논 문 제 목	**다속성 빅데이터로부터 유용한 정보 추출에 관한 연구 : 서울시 1인가구를 중심으로**
연 구 진	최정민 (건국대학교 건축대학 주거환경전공 교수, jmchoi@konkuk.ac.kr) 김건우 (건국대학교 일반대학원 건축공학과 석사과정)
공개자료 활용목록	서울시 서울서베이 도시정책지표조사 자료 (2010) 통계청 집계구별 인구·가구·주택 통계 (2010)

1 연구 배경 및 목적

○ 빅데이터가 증가하고 이용 가능해짐에 따라 대용량의 데이터에 숨어있는 흥미로운 규칙이나 패턴을 추출하여, 이를 새로운 정보로 활용하는 기술활용에 대한 관심이 점증하고 있음.

○ 본 연구는 이러한 배경을 바탕으로 대량의 다속성 범주형 자료 특성에 내재된 정보를 추출하기 위한 새로운 분석기법을 제안하고, 제안 기법을 적용하여 서울시 1인가구의 다양한 특성을 추출, 그 함의를 고찰함.

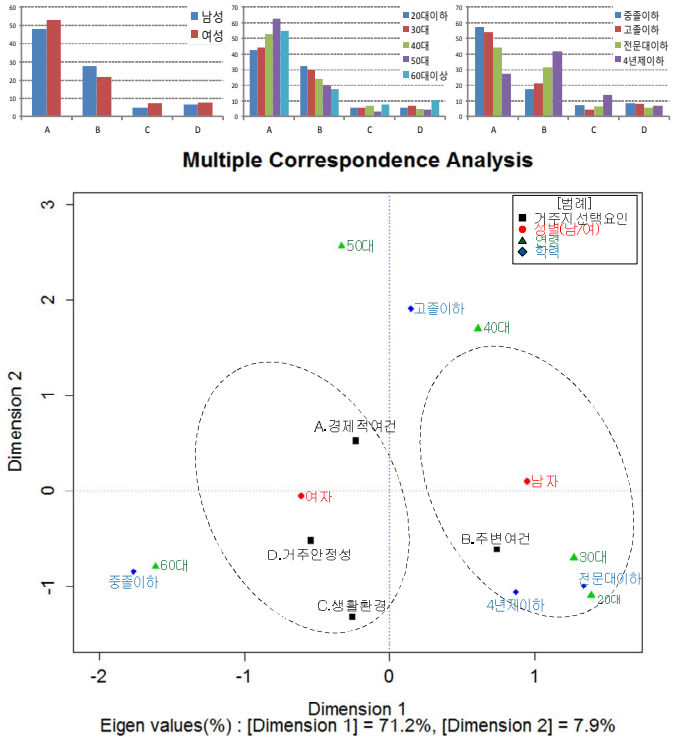
2 연구 주요내용

○ AIC지표에 의한 모형 선정과 대응분석에 의한 변수간의 상관관계를 알아보는 일련의 정보요약 과정을 제시함.

○ 첫째, 속성정보에 대한 데이터마이닝 기법을 서울서베이 2010 중 1인가구 2,133명을 대상으로 한 자료에 적용하여, 1인가구의 거주지 선택시 주요 고려사항을 그들의 인구통계학적 속성으로 대응관계를 고찰함.

○ 이 과정에서 8개의 이산형변수를 추출하여 AIC지표 제안수법의 유효성을 검증하고, 적합도 상위 6개 변수와 인구통계학적 변수를 다중대응분석(MCA)에 의해 상호 관계를 규명함.

○ 둘째, 공간정보에 대한 데이터마이닝 기법을 서울시 집계구별통계자료에 적용하여, 자치구별 1인가구 점유비율과 연속형 변수간의 상관분석을 실시하여 일정 상관관계 이상의 변수를 추출하여 이를 다차원척도법에 의한 배치관계 및 군집분석을 통해 유사한 지역을 유형화함.



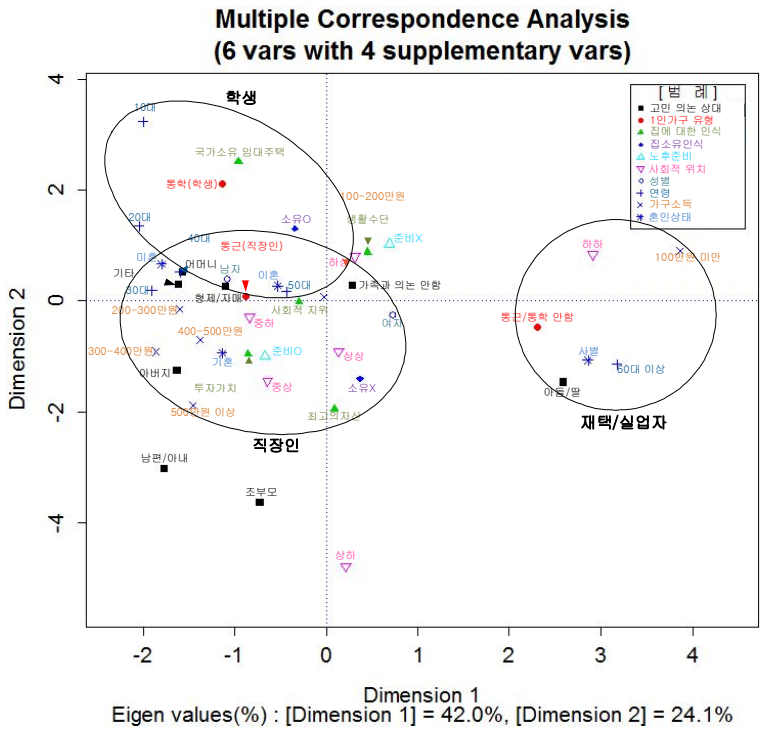
〈그림〉 1인가구 거주지 선택시 중요 고려사항과 인구통계학적 특성과의 관계

〈표〉 8개 유의한 변수 조합의 카이제곱 통계량

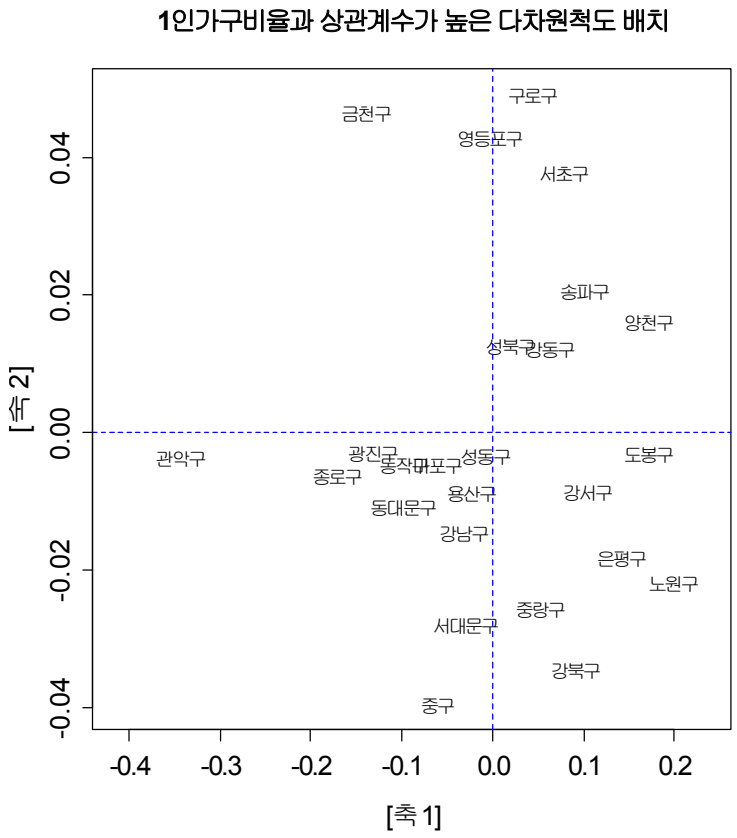
ID	변수 조합	카이값	유의수준
C1	q18xq32	571.8508	5.17056063396605e-113
C2	q36xq37	271.4976	1.51694898801563e-57
C3	q20xq32	240.7166	4.84509885803233e-46
C4	q15xq20	211.5883	9.40531775697842e-44
C5	q18xq20	208.9554	1.20475313049851e-26
C6	q15xq32	106.1459	8.92707663915856e-24
C7	q2xq32	74.7781	4.26394205560703e-14
C8	q20xq36	105.2465	1.43338284522869e-13
C9	q18xq37	72.4205	4.78363287113957e-13
C10	q32xq37	56.662	4.96590925118309e-13
C11	q2xq31	84.603	5.42552815519036e-13
C12	q15xq37	47.2688	6.18890303340976e-12
C13	q31xq36	83.6913	3.5723548974766e-11
C14	q15xq36	45.3908	3.2974906025061e-09
C15	q2xq18	71.3091	2.16412419097417e-07
C16	q18xq36	82.451	2.86405986015825e-07
C17	q31xq37	34.9577	4.73924755941177e-07
C18	q20xq37	34.385	1.99565780588922e-06
C19	q2xq15	27.9098	3.79394010945042e-06
C20	q2xq20	52.6378	4.42894353528747e-06
C21	q15xq18	32.3771	3.45719284835786e-05
C22	q20xq31	51.6657	0.0001273589412615 22
C23	q32xq36	26.9542	0.0007199898382745 46
C24	q2xq36	28.6863	0.0043849499037511 4
C25	q15xq31	14.3875	0.0061555904934249 2
C26	q31xq32	16.2081	0.0394970947081859

〈표〉 AIC지표에 의한 통계량 비교

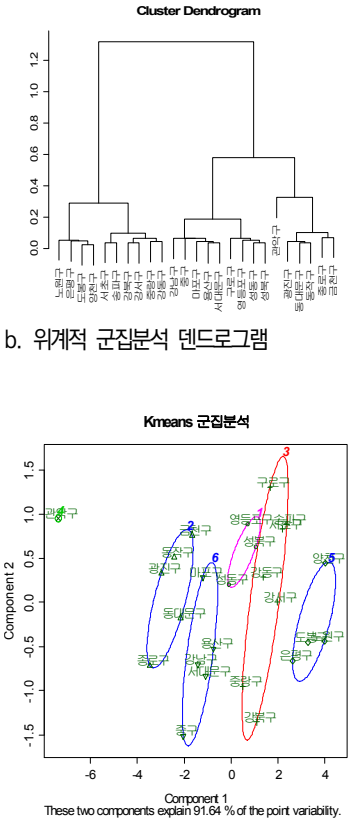
ID	변수 조합	AIC.I (①독립)	AIC.D (②비독립)	AIC차이 (①-②)
C1	q18xq32	9012.6057	8452.0036	-560.6021
C2	q36xq37	8672.7697	8393.3941	-279.3756
C4	q15xq20	8380.6232	8169.3686	-211.2546
C3	q20xq32	8561.1621	8362.289	-198.8731
C5	q18xq20	11287.0103	11176.7554	-110.2549
C6	q15xq32	6096.5958	5987.3147	-109.2812
C7	q2xq32	8011.2128	7945.0151	-66.1977
C8	q20xq36	11131.1623	11065.9296	-65.2327
C9	q18xq37	8827.1921	8767.2204	-59.9718
C11	q2xq31	9944.1281	9886.9033	-57.2248
C10	q32xq37	6090.5262	6037.5363	-52.9898
C13	q31xq36	10781.4678	10733.0813	-48.3866
C12	q15xq37	5911.9072	5865.8529	-46.0542
C14	q15xq36	8682.0717	8644.4284	-37.6433
C15	q2xq18	10746.4124	10716.0065	-30.4059
C16	q18xq36	11585.3367	11555.4878	-29.8489
C17	q31xq37	8025.0301	7997.854	-27.176
C18	q20xq37	8374.615	8349.5347	-25.0802
C20	q2xq20	10282.8739	10260.7945	-22.0794
C19	q2xq15	7832.5281	7810.5034	-22.0247
C21	q15xq18	8832.1322	8813.4209	-18.7113
C23	q32xq36	8861.8845	8849.5777	-12.3068
C22	q20xq31	10483.8732	10473.8506	-10.0226
C25	q15xq31	8029.6282	8023.203	-6.4252
C24	q2xq36	10597.7899	10593.4917	-4.2983
C26	q31xq32	8211.6201	8212.009	0.3889



〈그림〉 관심 이산형 변수 및 인구통계학적 변수를 투입한 다중대응분석 결과



a. 다차원척도법에 의한 자치구별 위상 관계



b. 위계적 군집분석 덴드로그램

c. K평균 군집분석에 의한 유형화

〈그림〉 다차원척도법 및 군집분석에 의한 자치구의 군집 및 유형화

3 연구 결과 및 기대효과(정책제언)

- ○ 1인가구에게는 경제, 사회, 교육환경보다는 주거환경이 우선적으로 마련되어야 주거안정이 확보됨
- ○ 거주지 선택에서도 1인가구의 성별 차이점이 발견되며, 연령과도 관계가 깊음
- ○ AIC지표에 의한 모형 선택과 대응분석으로 이어지는 정보요약 과정은 다속성 빅데이터에 매우 효과적이라는 것을 검증함
- ○ 1인가구는 6개의 자치구 군집으로 분류며, 군집별 특성은 1인가구의 인구통계학적 특성과 주거특성이 결합되어 도시 공간구조로 분화되어 나타남

4 공개자료 활용내용

- ○ 서울서베이 2010의 인구통계학적 속성 및 주거형태
- ○ 집계구별 인구, 가구, 주택 통계의 통계수치를 통한 분석테이블 작성