Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1201 ~1208

Environmental Survey Data Analysis by Data Fusion Techniques

Kwang-Hyun Cho[1] · Hee-Chang Park[2]

Abstract

Data fusion is generally defined as the use of techniques that combine data from multiple sources and gather that information in order to achieve inferences. Data fusion is also called data combination or data matching. Data fusion is divided in five branch types which are exact matching, judgemental matching, probability matching, statistical matching, and data linking.

Currently, Gyeongnam province is executing the social survey every year with the provincials. But, they have the limit of the analysis as execute the different survey to 3 year cycles. In this paper, we study to data fusion of environmental survey data using sas macro. We can use data fusion outputs in environmental preservation and environmental improvement.

Keywords : Data fusion, Data combination, Data matching, Macro, SAS, Social survey

1. 서론

사회지표조사는 변화하는 역사적 흐름 속에서 우리가 처해 있는 사회적 상태를 종합적이고 집약적으로 나타냄으로써 사회구성원들의 삶의 질을 전반적으로 파악하고 사회변화를 포착할 수 있는 조사이다. 그동안 환경부문의 사회지표조사 데이터에 대해 자료의 수집과 기초적인 분석방법, 다변량 분석방법 등에 중점을 두고 연구가 활발히 이루어지고 있다(이상훈, 1995, 문상기와 우남칠, 2001, 김정태 등, 2003).

현재 경상남도는 경상남도 도민들을 대상으로 매년 환경, 교통 등의 부문에 대하여

1) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : cho1023@changwon.ac.kr

2) Corresponding Author : Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : hcpark@changwon.ac.kr

사회지표 조사를 실시하고 있다. 그러나 사회지표 조사 자료에 대하여 3년 주기로 매년 설문 문항을 다르게 하여 설문조사를 실시하고 있어 도민들의 환경의식에 대한 분석 시 연도별로 각각 분석을 실시해야 함으로서 유기적인 분석이 가능하지 못한 실정이다. 또한, 특정 연도의 사회지표 조사 자료에서는 환경의식 분석에 사용할 환경 관련 문항들이 기타 연도에 비하여 작아 다양한 분석을 실시하지 못하고 있어 분석의 한계점이 있다. 이에 각 연도의 사회지표 조사 자료의 환경 관련 문항을 결합하여 하나의 데이터 파일을 만들면 부가성이 높은 정보를 획득할 수 있을 것이며 이를 위하여 사용되어 지는 방법 중의 하나가 데이터 퓨전(data fusion) 방법이다.

데이터 퓨전은 같은 모집단에서 나온 서로 다른 표본들을 포함하는 데이터 셋을 합치는 기법 또는 처리과정으로 정의되며 데이터 융합, 데이터 결합, 데이터 매칭이라고 불리기도 한다. 데이터 퓨전은 그 자체가 하나의 분석이며 최종 결과라기보다는 통계분석 결과의 질을 높이기 위한 방법이라고 할 수 있다. 즉, 데이터 퓨전을 통해서 얻은 최종 결과에 대한 추가된 정보를 이용함으로써 통계 분석의 질을 향상시킬 수 있다. 데이터 퓨전은 1970년대부터 미국과 독일 등에서 주로 경제 통계학 분야에서 적용해 왔다. 데이터 퓨전은 1980년대에 들어와서 점차 일반화되었고, 특히 미디어연구와 미시경제 분석에서 널리 활용되어 왔으며 현재 다양한 분야에서 적용되어 지고 있다.

데이터 퓨전에 대한 국내 연구로는 최기주와 정연식(1998)은 교통정보의 생성을 위하여 링크 통행시간 추정을 하는데 데이터 퓨전을 이용하였고 손소영과 이성호(2000)은 도로교통사고 자료처리에서의 교통사고 심각도 분류분석에 데이터 퓨전을 적용하였으며 신형원과 손소용(2000)은 다구찌 디자인을 이용한 데이터 퓨전의 성능 비교에 대하여 연구한바 있다. 또한 박성원 등(2001)은 데이터 퓨전을 이용하여 얼굴영상 인식 및 인증에 관하여 연구한바 있으며 김호종 등(2003)은 신경망 기반 추천 모델의 성능향상을 위하여 데이터 퓨전을 적용하였다.

본 논문에서는 경상남도에서 2001년과 2002년에 조사된 사회지표 조사 자료의 환경 관련 문항에 대하여 데이터 퓨전을 적용하여 이를 분석하고자 한다. 데이터 퓨전은 정확 결합, 판단 결합, 확률적 결합, 통계적 결합, 데이터 연결의 5가지 종류로 구분되며, 사회지표 조사 자료에 대해서는 공통으로 가지는 변수에 개인 식별 가능한 변수가 없기 때문에 통계적 결합 방법을 사용하여 퓨전을 실시한다. 2001년과 2002년에 조사된 사회지표 조사 자료의 데이터 퓨전으로 각각의 데이터가 결합하여 하나의 데이터 파일로 생성되므로 자료의 정보가 보다 풍부해진다. 이는 각각의 자료에서는 불가능한 통계 분석이 가능해 질 수 있을 뿐만 아니라 도민들의 환경의식을 더욱더 총체적으로 분석할 수 있는 기초 자료를 제공할 수 있다.

본 논문의 2절에서는 데이터 퓨전에 대하여 기술하고 3절에서는 데이터 퓨전에 의한 사회지표조사 분석에 대하여 기술하며, 4절에서 결론을 맺는다.

2. 데이터 퓨전

데이터 퓨전은 같은 모집단에서 나온 서로 다른 표본들을 포함하는 데이터 셋을 합치는 기법 또는 처리 과정으로 정의된다. 데이터 퓨전은 별개의 데이터 파일을 결합하여 하나의 완전한 데이터 파일을 만드는 것을 의미하는 것으로 데이터 융합, 데이터 결합, 데이터 매칭이라고 불리기도 한다(한상훈 등(2004)).

데이터 퓨전은 그 자체가 하나의 분석이며 최종 결과라기보다는 통계분석 결과의 질을 높이기 위한 방법이라고 할 수 있다. 즉, 데이터 퓨전을 통해서 얻은 최종 결과에 대한 추가된 정보를 이용함으로써 통계 분석의 질을 향상시킬 수 있다.

서로 다른 두 개의 파일 A와 B를 가정하자. 파일 A와 B에는 X라는 변수가 공통적으로 존재하고 Y변수와 Z변수는 파일 A와 파일 B에 각각 존재한다고 하자. 즉, 파일 A와 B에 공통적으로 X라는 변수가 있고 파일 A에는 Y라는 변수만 존재하며 파일 B에는 Z라는 변수만 존재한다고 하자. 변수 X, Y, Z로 구성된 파일을 만들기 위하여 파일 A와 B를 결합하여 하나의 파일로 만들면 된다. 예를 들어, 파일 A에는 변수 X와 변수 Y로 구성되어 있고 파일 B에는 변수 X와 변수 Z로 구성되어 있다고 하자. 여기서, 파일 A와 파일 B에 공통적으로 존재하는 변수 X를 공통 변수라고 하고 파일 A 또는 파일 B에서만 존재하는 변수 Y와 변수 Z를 고유 변수라고 한다. 데이터 퓨전의 결과로 생성된 결합 파일은 두 파일의 공통변수 X를 이용하여 파일 B에 존재하는 변수 Z를 파일 A에 추가한 형식으로 나타난다. 여기서, 변수 Z를 수용하는 파일 A를 수용 파일이라 하고 변수 Z를 제공하는 파일 B를 제공 파일이라고 한다. 파일 A와 B에 의하여 데이터 퓨전을 수행한 후 생성된 파일을 결합 파일이라고 한다. 영국 National Statistics(2003)에 따르면 데이터 결합(Data Matching)의 종류는 정확 결합(Exact Matching), 판단 결합(Judgemental Matching), 확률적 결합(Probability Matching), 통계적 결합(Statistical Matching), 데이터 연결(Data Linking)로 구분된다.

본 논문에서 사용한 통계적 결합은 공통으로 가지는 변수는 존재하나 고유 식별 변수처럼 개인 식별 가능한 변수가 없을 때 회귀분석, 로지스틱 회귀분석 등의 통계적 방법을 사용하여 데이터를 결합하는 방법이다. 한상훈 등(2004)에 의하여 연구되어진 통계적 결합에 대한 알고리즘은 결합하고자 하는 변수가 범주형(이분형)인 경우와 연속형인 경우로 구분된다. 결합하고자 하는 변수가 범주형(이분형)인 경우에는 로지스틱 회귀분석을 이용하여 자료를 결합하고, 연속형인 경우에는 회귀분석을 이용하여 결합을 실시한다.

3. 데이터 퓨전에 의한 사회지표조사 분석

3.1 사회지표조사 자료

2001년과 2002년 조사된 사회지표 조사에 대한 데이터 퓨전 자료는 <표 1> 및 <표 2>와 같다.

<표 1> 2001년 사회지표 조사 자료 구조

변수 구분	변수명	변수형
고유변수	주관적 상수도 오염도	연속형
	주관적 하수도 오염도	연속형
	주관적 소음진동 오염도	연속형
	주관적 악취 오염도	연속형
	주관적 대기 오염도	연속형
	주관적 토양 오염도	연속형
공통변수	연령	연속형
	주관적 사회계층	연속형
	학력	범주형
	성별	범주형
	결혼유무	범주형

<표 2> 2002년 사회지표 조사 자료 구조

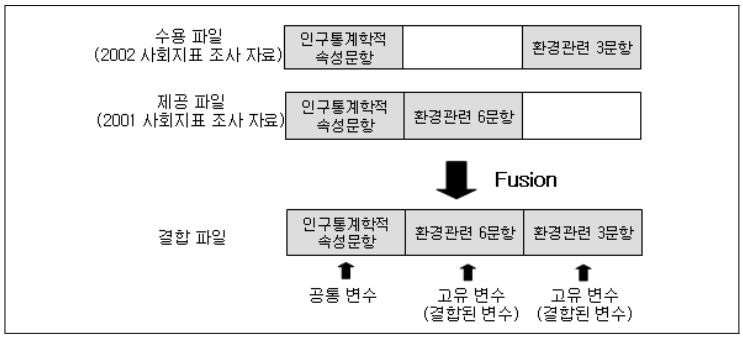
변수 구분	변수명	변수형
고유변수	쓰레기분리 수거의 참여 정도	범주형
	녹색제품의 구입 여부	범주형
	수돗물 음용수 적정 여부	범주형
공통변수	연령	연속형
	주관적 사회계층	연속형
	학력	범주형
	성별	범주형
	결혼유무	범주형

각각 데이터는 총 9,999건과 9,877건이며 2001년 사회지표 조사에서는 고유변수 6문항과 공통변수 5문항으로 구성되어 있고 2002년 사회지표 조사에서는 고유변수 3문항과 2001년 사회지표 조사와 동일한 공통변수 5문항으로 구성되어 있다.

3.2 데이터 퓨전

2001년 조사된 사회지표 조사와 2002년 조사된 사회지표 조사에 대한 환경 관련 문

항의 데이터 퓨전 방법은 <그림 1>과 같다. 각 조사 자료에 대해서는 공통으로 가지는 변수에 개인 식별 가능한 변수가 없기 때문에 통계적 결합 방법을 사용하여 퓨전을 실시한다. 수용파일은 2002년 사회지표 조사 자료이고 제공파일은 2001년 사회지표 조사 자료가 된다.



<그림 1> 사회지표 조사의 데이터 퓨전

본 논문에서는 박희창과 조광현(2006)에 의해 연구되어진 통계적 결합에 대한 SAS 매크로 프로그램을 사용하여 사회지표조사 자료를 결합하였다. 데이터 퓨전 결과, 총 데이터의 수는 19,876건이고 변수는 공통변수 5문항, 2001년 사회지표조사의 고유변수 6문항, 2002년 사회지표조사의 고유변수 3문항의 총 14문항으로 구성되어 있다.

3.3 분석 결과

사회지표조사 자료는 3년 주기로 매년 설문 문항을 다르게 하여 설문조사를 실시하고 있어 각 연도 마다 분석을 각각 실시해야 하며, 각 연도별 사회지표조사의 환경관련 문항을 통합적으로 분석을 할 수 없는 문제점이 있다. 그러나 데이터 퓨전에 의한 사회지표조사를 결합하면 다음과 같은 분석이 가능해진다.

첫째, 각 연도별 환경관련 문항의 응답 결과를 비교 분석할 수 있다.
둘째, 각 연도별 환경관련 문항의 응답 결과를 종합적으로 분석할 수 있다.

각 연도별 환경관련 문항의 응답 결과의 비교는 <표 3> 및 <표 4>와 같다. <표 3>은 연속형 문항에 대한 t-검정 결과이다. <표 3>에서 보는 바와 같이 연속형 환경관련 6개 문항에 대하여 모두 통계적으로 차이(유의수준 : 0.05)가 없는 것으로 나타나, 2001년과 2002년에 대한 주관적 환경 오염도의 응답은 차이가 없다고 할 수 있다.

<표 3> 연속형 문항에 대한 분석 결과

문항	구분	평균	표준편차	유의확률
주관적 상수도 오염도	2001년 자료	3.13	0.833	0.395
	2002년 자료	3.14	0.379	
주관적 하수도 오염도	2001년 자료	2.99	0.834	0.738
	2002년 자료	2.99	0.362	
주관적 소음진동 오염도	2001년 자료	2.96	0.834	0.956
	2002년 자료	2.96	0.632	
주관적 악취 오염도	2001년 자료	3.03	0.867	0.355
	2002년 자료	3.04	0.383	
주관적 대기 오염도	2001년 자료	3.24	0.873	0.879
	2002년 자료	3.25	0.390	
주관적 토양 오염도	2001년 자료	3.30	0.759	0.829
	2002년 자료	3.31	0.337	

<표 4> 범주형 문항에 대한 교차분석 결과

쓰레기분리 수거의 참여 정도			잘 참여	잘 참여하지 않음	유의확률
구분	2001년 자료	빈도(%)	7101	2898	0.006^{***}
		수정된 잔차	-2.7	2.7	
	2002년 자료	빈도(%)	6762	2528	
		수정된 잔차	2.7	-2.7	
녹색제품의 구입 여부			구매	구매 않음	유의확률
구분	2001년 자료	빈도(%)	3233	6766	0.000^{***}
		수정된 잔차	-12.3	12.3	
	2002년 자료	빈도(%)	4020	5855	
		수정된 잔차	12.3	-12.3	
수돗물 음용수 적정 여부			적당함	적당하지 않음	유의확률
구분	2001년 자료	빈도(%)	2188	7811	0.000^{***}
		수정된 잔차	-8.4	8.4	
	2002년 자료	빈도(%)	2348	6297	
		수정된 잔차	8.4	-8.4	

* : 유의수준 0.1, ** : 유의수준 0.05, *** : 유의수준 0.01

<표 4>는 범주형 문항에 대한 교차분석 결과이다. <표 4>에서 보는 바와 같이 모든 문항에 대하여 유의한 차가 있는 것으로 나타났다. 쓰레기분리 수거의 참여 정도 문항에서는 2001년도에는 2002년도에 비하여 잘 참여하지 않음의 응답비율이 높은 반면, 2002년도에서는 잘 참여의 응답 비율이 높은 것으로 나타났다. 녹색제품의 구입여부 정도 문항에서는 2001년도에는 2002년도에 비하여 구매 않음의 응답비율이 높은 반면, 2002년도에서는 구매의 응답 비율이 높은 것으로 나타났다. 수돗물 음용수 적정 여부 문항에서는 2001년도에는 2002년도에 비하여 적당하지 않음의 응답비율이 높은 반면, 2002년도에서는 적당함의 응답 비율이 높은 것으로 나타났다.

데이터 퓨전에 의하여 통합된 사회지표조사 자료에 대한 종합적 분석을 실시하기 위하여 인구통계학적 문항과 환경관련 문항에 대하여 집단간 차의 검정(t-test, 분산분석, 교차분석)을 실시하였다. 분석 결과, <표 5>와 같이 성별에 대하여 환경관련문항 간의 차이가 있음을 알 수 있었다. <표 5>에서는 통계적으로 유의한 결과만 제시

하였다.

<표 5> 성별과 환경관련문항간의 분석 결과

문항	구분	평균	표준편차	유의확률
주관적 하수도 오염도	남성	2.98	0.635	0.007***
	여성	3.00	0.651	
주관적 소음진동 오염도	남성	2.95	0.722	0.037**
	여성	2.97	0.753	
주관적 대기 오염도	남성	3.23	0.671	0.000***
	여성	3.26	0.683	
주관적 토양 오염도	남성	3.27	0.587	0.000***
	여성	3.31	0.588	

* : 유의수준 0.1, ** : 유의수준 0.05, *** : 유의수준 0.01

<표 5>에서 보는 바와 같이 주관적 하수도 오염도, 주관적 소음진동 오염도, 주관적 대기 오염도, 주관적 토양 오염도 문항에 대하여 통계적으로 유의한 차가 있는 것으로 나타났다. 이를 자세히 살펴보면 모든 문항에 대하여 여성이 남성보다 주관적 환경 오염도에 대하여 긍정적인 응답을 보이고 있는 것을 알 수 있다.

4. 결론

현재 경상남도는 경상남도 도민들을 대상으로 매년 환경, 교통 등의 부문에 대하여 사회지표 조사를 실시하고 있다. 그러나 3년 주기로 매년 설문 문항을 다르게 하여 설문조사를 실시하고 있어 도민들의 환경의식에 대한 분석 시 연도별로 각각 분석을 실시해야 함으로서 유기적인 분석이 가능하지 못하여 분석의 한계점이 있다. 이에 본 논문에서는 각각의 자료를 효율적으로 사용하기 위하여 경상남도에서 2001년과 2002년에 조사된 사회지표 조사 자료의 환경관련 문항에 대하여 SAS 매크로를 이용하여 데이터 퓨전을 실시하였다. 데이터 퓨전 결과 2001년과 2002년에 조사된 사회지표 조사 자료의 데이터가 결합하여 하나의 데이터 파일로 생성되어 각각의 자료에서는 불가능한 통계 분석이 가능해 질 수 있을 뿐만 아니라 도민들의 환경의식을 더욱더 총체적으로 분석할 수 있는 기초 자료를 제공할 수 있다. 향후 연구 과제로 데이터 퓨전의 나머지 방법의 SAS 매크로 적용 방안에 대한 연구가 필요하며 데이터 퓨전 결과의 유효성 평가 방법에 대한 연구가 필요할 것이다.

참고 문헌

1. 김정태, 정진도, 김광석 (2003). 여름철 충청남도 서북부 지역에서의 대기오염물질 농도 분포특성에 관한 연구, *대한환경공학회 2003 춘계학술발표회 논문집*, pp.1326-1328.
2. 김호종, 김은주, 김명원 (2003). 신경망 기반 추천 모델의 성능향상을 위한 정보의 융합, *한국정보과학회 학술발표논문집*, Vol. 2003, No. 2483,

pp.422-424.
3. 문상기, 우남칠 (2001). 통계분석을 이용한 지하수위 변동 특성 분류, *한국지하수토양환경학회 01 추계학술발표회논문집*, Vol. 2001, pp.155-159
4. 박성원, 권지웅, 최진영 (2001). 데이터 퓨전을 이용한 얼굴영상 인식 및 인증에 관한 연구, *퍼지 및 지능시스템학회 논문집*, Vol. 11, No. 4, pp.302-306.
5. 박희창, 조광현 (2006). 통계적 데이터 퓨전을 위한 sas 매크로, *한국자료분석학회 추계학술대회*, pp.142-151.
6. 신형원, 손소영 (2000). 다구찌 디자인을 이용한 데이터 퓨전 및 군집분석 분류 성능 비교, *대한산업공학회/한국경영과학회 2000년 춘계공동학술대회 논문집*, Vol. 2000, pp.601-604.
7. 손소영, 이성호 (2000). 데이터 융합, 앙상블과 클러스터링을 이용한 교통사고 심각도 분류분석, *대한산업공학회/한국경영과학회 2000년 춘계공동학술대회 논문집*, Vol. 2000, pp.597-600.
8. 이상훈 (1995). 수질자료의 추세분석을 위한 비모수적 통계검정에 관한 연구, *환경영향평가*, Vol. 4, No. 2, pp.93-103
9. 최기주, 정연식 (1998). 링크 통행시간 추정을 위한 데이터 퓨전 알고리즘의 개발, *대한교통학회지*, Vol. 16, No. 2, pp.177-195.
10, 한상훈, 안일호, 하덕주, 최종후 (2004). 데이터 퓨전과 평가, *한국데이터마이닝학회 2004 추계학술대회*, pp.238-254.
11. National Statistics (2003). National Statistics Code of Practice Protocol on Data Matching. http://www.statistics.gov.uk/about/consultations/general_consultations /downloads/ Protocol_on_Data_Matching.pdf

[2006년 10월 접수, 2006년 11월 채택]