통계연구(2009), 제14권 제1호, 82-98

통계조사자료와 행정자료 간의 통계적 매칭기법에 관한 연구

이영섭[1] · 김선웅[2] · 안홍엽[3] · 임경은[4] · 김희경[5]

요약

통계적 매칭에 의한 데이터 통합이란 보유하고 있는 데이터 파일에 필요한 변수가 없거나 결측값이 존재할 경우, 다른 원천 데이터로부터 모아진 데이터와 정보를 통합하는 것이다. 이와 같은 데이터 통합을 통해 데이터의 질을 상당히 높일 수 있을 뿐만 아니라, 다른 조사를 통해 데이터를 새로 얻는 것보다 시간과 비용을 절약하고 응답자의 부담도 줄일 수 있다. 즉, 효율적인 분석을 가능하게 하는 새로운 통합 데이터를 만드는 것이 통계적 매칭 기법의 목적이다. 본 연구에서는 통계조사자료와 행정자료를 통합하여 양질의 데이터를 생성하기 위한 효율적인 통계적 매칭기법 및 평가방법에 대한 연구를 수행한다. 또한 사례연구를 통해 랜덤 핫덱 방법을 이용한 국민연금자료와 사업체 기초조사자료의 통계적 매칭을 수행하고, 매칭 결과를 평가한다. 이러한 결과는 향후 행정자료의 효율적 활용을 도모하는 데 매우 중요한 역할을 할 것이다.

주요용어: 데이터 통합, 통계적 매칭, 수용파일, 제공파일, 랜덤 핫덱 방법

1. 서론

공공기관이나 기업이 효과적인 데이터 분석을 하기 위해서는 조사단위인 개인이나 가구들에 대한 기본적인 인구통계학적 특성, 취미와 생활 습관, 기호 등의 다양한 정보를 얻은 후 접근해야 한다. 그러나 현재 대부분의 공공기관이나 기업들이 보유한 데이터에서는 조사단위에 대한 충분한 설명을 확보하거나 이에 접근하는 일이 어려운 경우가 많다. 또한 원천 데이터 소스의 다양성, 단일 데이터의 불충분성, 부서 간의 데이터 공유의 부족으로 인하여 하나의 데이터에서 분석에 필요한 모든 정보를 얻는다는 것은 매우 어려운 일이다.

이러한 문제는 데이터 매칭(data matching) 또는 데이터 통합(data fusion)을 통해 많은 부분 보완할 수 있다. 일반적인 조사 데이터에는 대체로 나이, 성별 등 공통적으로 포함하고 있는 사항들이 몇 가지 있다. 이러한 공통 요소들을 기본으로 완전히 같지는 않지만 비슷한 사람이나 집단끼리의 정보는 얻을 수 있을 것이다.

본 연구에서는 통계조사자료와 행정자료를 효율적으로 매칭시킬 수 있는 통계적 기법에 대한 연구를 수행하고자 한다. 통계조사자료와 행정자료 간 매칭 및 분석을 통하여 신규 통계의 작성을 가능하게 하고, 재구성된 새로운 데이터를 미래 연구에 사용 가능한지를 확인한다. 또한 다양한 목적에 따라 수집된 데이터들 간의 효율적인

1) 동국대학교 통계학과, 부교수. E-mail: yung@dongguk.edu
2) 동국대학교 통계학과, 부교수. E-mail: sunwk@dongguk.edu
3) 동국대학교 통계학과, 조교수. E-mail: ahn@dongguk.edu
4) 교신저자. 통계청 통계개발원, 통계사무관. E-mail: kelim06@nso.go.kr
5) 동국대학교 대학원 통계학과, 박사과정. E-mail: khk0228@dongguk.edu

통합을 위한 통계적 매칭 기법에 대하여 알아보고, 통계조사자료와 행정자료를 통합하여 양질의 데이터를 생성하기 위한 새로운 매칭 기법을 개발하는 것이 본 연구의 목적이다.

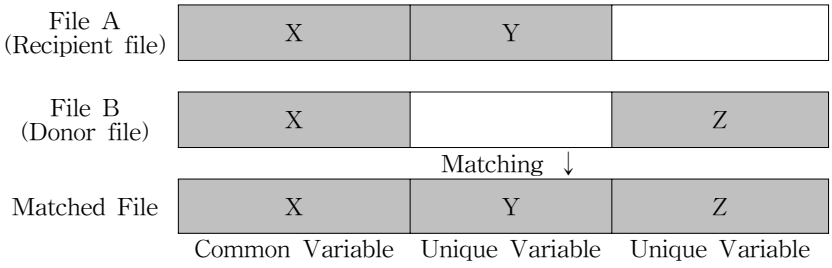
2. 데이터 매칭

우리가 통계분석을 수행할 때, 필요로 하는 변수를 모두 포함하는 데이터 파일은 흔하지 않기 때문에 여러 데이터 파일을 이용해서 필요한 변수를 매칭시켜 사용한다. 매칭을 통한 방법은 다른 조사를 통해서 데이터를 얻는 것보다 시간과 비용을 절약할 수 있고 때로는 분석과 추정에 있어서 더욱 신뢰성을 높일 수 있으며, 조사 응답자의 부담을 줄여줄 수 있다.

데이터 매칭은 별개의 데이터 파일을 결합하여 하나의 데이터 파일을 만드는 방법으로 영국의 "National Statistics code of Practice Protocol on Data Matching(2003)"에 따르면 정확 매칭(Exact Matching), 판단 매칭(Judgemental Matching), 확률적 매칭(Probability Matching), 통계적 매칭(Statistical Matching), 데이터 연결(Data Linking) 등 크게 5가지의 종류로 나누어 볼 수 있다.

먼저 정확 매칭이란 주민등록번호, 국가보험번호, 사회보장번호와 같이 ID를 나타낼 수 있는 변수가 공통으로 있는 경우와 같이 변수 값이 완전히 일치하는 경우에 데이터를 결합하는 방법이다. 같은 사람 또는 같은 물건을 완벽하게 결합할 수 있고, 공통 변수에 측정오차가 없다면 이상적으로 데이터 매칭을 수행할 수 있는 장점이 있다. 반면 사람과 관련된 경우에 개인의 고유한 정보를 이용해야 하므로 사생활 침해의 여지가 있다는 단점이 있다. 판단 매칭이란 공통인 변수들 사이에 정확히 일치하는 것은 없지만 데이터에 대해 잘 알고 있는 경우, 또는 몇 가지 조사를 시행하고 적절하다고 판단하는 것을 결합하는 방법이다. 확률적 매칭은 정확 결합의 경우에서 공통 변수들에 오류가 있는 경우에 정확한 정도에 따라 가중치를 주고 확률적으로 데이터를 결합하는 방법이며, 통계적 매칭이란 공통으로 가지는 변수에 개인 식별 가능한 변수가 없을 때 수행하는 데이터 결합 방법이다. 마지막으로 데이터 연결이란 둘 이상의 파일에서 변수들 간의 연관성을 만들어 내어 바로 데이터 갱신이 가능하도록 하는 데이터 결합 방법이다.

<그림 2.1>에 나타나있는 바와 같이 서로 다른 경로로 얻어진 두 개의 파일을 고려해 보자. File A는 (X Y)로 구성되어 있고 File B는 (X Z)로 구성되어 있다고 하자. File A와 File B에 모두 관찰되는 변수 X를 공통변수(common variable)라 하고 File A에서만 관찰되는 변수 Y와 File B에서만 관찰되는 변수 Z를 유일변수(unique variable)라고 한다. 일반적으로 데이터 매칭을 수행하면 공통변수를 이용하여 File B에 있는 Z를 File A에 추가하게 된다. 이때, File A를 수용파일(recipient file)이라 하고 File B를 제공파일(donor file)이라고 하며, 데이터 매칭을 수행한 후 생성된 파일을 결합파일(matched file)이라 한다.



<그림 2.1> 데이터 매칭

2.1 통계적 매칭

가. 통계적 매칭의 구분

통계적 매칭(Statistical Matching)을 수행할 때, 접근방법을 두 가지로 나누어 볼 수 있다. 먼저, 수용파일에서 관찰되지 않은 변수를 예측하는데, 특정모형을 가정하지 않고 전적으로 데이터에 기초해서 통계적 결합을 수행하는 접근 방법으로 사전 준비 작업이 거의 없고 수행하기 쉽다는 장점이 있다. 반면 계산 시간이 오래 걸린다는 단점이 있다.

다른 접근 방법으로는 데이터의 특징을 잘 반영하는 모형을 사용하여 접근하는 방법으로 추상적인 모형을 만들어 관찰되지 않은 값을 예측하게 된다. 이 경우 모형을 가정하지 않는 경우에 비해 일반화하기 쉽다는 장점은 있으나, 데이터의 크기가 매우 큰 경우나 데이터가 모형에 대한 가정에 맞지 않는 경우에는 적용이 어렵다.

또한 통계적 매칭을 수행하는 방법에 따라 제약이 있는 매칭(constrained matching)과 제약이 없는 매칭(unconstrained matching)으로 구분된다. 제약이 없는 매칭은 수용파일에 있는 모든 개체가 매칭파일에서 나타나고, 제공파일의 모든 개체가 데이터매칭 과정에서 사용될 필요는 없다. 이러한 매칭은 매칭파일에서 Z변수의 주변분포가 원래의 제공파일에서의 분포와 달라질 수 있다는 단점이 있다.

제약이 있는 매칭은 수용파일과 제공파일에 있는 모든 개체들이 한번 이상 매칭과정에서 이용되며, 데이터매칭을 수행했을 때 두 파일에 있는 모든 개체들이 매칭파일에 나타난다. 이러한 매칭은 공통변수인 X변수들 사이의 연관 관계가 작아도 매칭이 된다는 단점이 있다.

나. 통계적 매칭의 제약조건

Van der Puttern et al. (2002)은 데이터 매칭이 유용한 결과를 도출하기 위해 다음과 같은 제약조건을 제시하였다. 첫째, 제공파일은 수용파일을 대표할 수 있어야 한다. 그러나 반드시 두 데이터가 같은 모집단에서 추출될 필요는 없다. 둘째, 공통변수 X가 주어졌을 때, 유일변수인 Y와 Z사이에 조건부 독립관계($P(Y,Z|X) = P(Y|X) \cdot P(Z|X)$)가 성립되어야 한다. 이러한 조건부 독립성(CIA ; conditional independent assumption)을 가정하는 이유는 수용파일과 제공파일 각각으로부터 X, Y, Z의 결합확률분포함수(joint probability distribution function) $f(x,y,z)$를 추정할

수 없기 때문이다.

만약 CIA가 만족된다면 (x,y,z)의 결합확률 분포함수는 다음과 같다.

$$f(x,y,z)=f(y|x)f(z|x)f(x)$$

여기서 $f(y|x)$는 수용파일로부터 추정 가능하고, $f(z|x)$는 제공파일로부터 추정 가능하다. 그러면 $f(x,y,z)$가 추정 가능하며 다음과 같이 $f(y,z)$도 추정 가능하다.

$$f(y,z)=\int_{-\infty}^{\infty} f(x,y,z)dx \quad \text{for continuous } x$$

이는 CIA가 만족되면 매칭 후 각각의 데이터로부터는 추정할 수 없었던 Y와 Z의 관계를 파악할 수 있게 된다는 것을 의미한다.

2.2 통계적 매칭 알고리즘

가. 단계적 매칭 알고리즘

Case 1. 결합하려는 변수가 범주형인 경우

로지스틱회귀분석의 결과를 이용하여 데이터의 근사성을 측정한다. 이때, 수용파일과 제공파일의 각 개체를 추정된 회귀식에 적합시켜 얻은 값을 근사성 측정을 위한 점수로 사용하게 된다.

$$D_{ij}^{F}=|\hat{Z}_{1i^{R}}-\hat{Z}_{1j^{D}}| \quad \text{for given } i$$

여기서 $\hat{Z}_{1i^{R}}$ 는 제공파일에서 추정된 회귀식을 수용파일에 적합시켜 구한 값이며, $\hat{Z}_{1j^{D}}$는 제공파일에 적합시켜 구한 값이다.

만약 측정한 근사성 정도가 같아지게 되면, 수용파일 하나의 개체에 여러 개의 제공파일 개체가 결합하게 된다. 이러한 경우 로지스틱회귀분석 결과 추정된 회귀식에 포함되지 않은 범주형 변수들을 이용하여 두 번째 근사성을 측정하여 이용하게 되며, 두 번째 단계에서 측정한 근사성도 같은 값을 갖게 되면 이용하지 않은 연속형 변수로 근사성을 측정한다. 최종적으로 근사성 측정값이 작은 값을 갖는 수용파일의 i개체와 제공파일의 j개체를 결합한다.

Case 2. 결합하려는 변수가 연속형인 경우

범주형 변수를 결합할 때 이용한 방법 그대로 적용한다. 단, 결합하고자 하는 연속형 변수를 종속변수로 하고 나머지 변수를 독립변수로 하는 선형회귀분석을 수행한다.

Case 3. 범주형 변수와 연속형 변수를 동시에 결합하는 경우

하나의 변수를 결합하는 것보다 여러 개의 변수를 한 번에 결합하는 경우가 더 일반적이다. 결합하려는 변수가 범주형인 경우와 연속형인 경우의 단계별 순위합으로 근사성을 측정한다.

$$D_{ij}^{RF} = RD_{ij}^{FZ_1} + RD_{ij}^{FZ_2} \quad \text{for given} \quad i$$

여기서 $RD_{ij}^{FZ_1}$는 범주형 변수인 Z_1를 결합할 때의 D_{ij}^{F} 순위이며, $RD_{ij}^{FZ_2}$는 연속형 변수인 Z_2를 결합할 때의 D_{ij}^{F} 순위이다.

이때 D_{ij}^{RF}값이 작은 수용파일의 i개체와 제공파일의 j개체를 결합하게 되며, 결합하려는 변수가 범주형인 경우와 연속형인 경우의 두 번째 단계의 순위합으로 근사성을 측정한다. 이때 근사성이 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다.

나. k-최근접이웃 매칭 알고리즘

최근접이웃방법은 통계적 매칭에 가장 흔히 사용되는 방법으로 가장 유사한 하나의 개체를 매칭에 사용하는 방법이다. 여기서 한 단계 나아가 상대적으로 유사한 k개의 개체를 선택하여 매칭에 사용하는 방법이 k-최근접이웃방법이다. 실제 사례로 D'Orazio et al (2006)의 Survey on Household Income and Wealth (SHIW) 데이터와 Household Budget Survey (HBS)의 데이터의 매칭 연구가 있다.

다. 회귀분석 매칭 알고리즘

회귀분석을 적용하여 매칭을 하는 방법은 먼저 하나의 데이터 파일에서 회귀모형을 추정한 후, 추정된 회귀모형을 이용하여 두 개의 데이터 파일에서 대한 예측치를 구하는 것이다. 그리고 두 파일의 예측치 사이의 거리가 가장 짧은 개체를 찾으면 매칭이 이루어진다.

최근접이웃 접근방법은 데이터 매칭이 이루어질 때 공통변수 X만을 이용하지만, 회귀분석 접근방법은 공통변수 X뿐만 아니라 제공파일의 유일변수 Z를 이용한다는 데 그 차이가 있다.

라. 회귀분석과 k-최근접이웃방법의 결합 매칭 알고리즘

회귀분석 방법을 이용한 통계적 매칭 방법은 추정치의 거리가 가장 가까운 하나의 개체만을 사용함으로써 상대적으로 유사한 다른 개체들의 정보를 무시하게 된다. 상대적으로 유사한 개체에 대한 정보손실을 줄여 데이터 통합기법의 성능을 높이고자 회귀분석기법에 k-최근접이웃 접근법을 결합하여 가장 가까운 하나의 개체가 아니라 k개의 개체를 이용하여 통합변수를 추가시키는 방법이다.

마. 랜덤 핫덱 방법

랜덤 핫덱(Random Hot Deck)은 수용자 파일의 각 관측치에 대해 제공자 파일의 관측치를 랜덤하게 선택하여 매칭시키는 방법이다. 특히 수용자 파일과 제공자 파일의 관측치들은 대개 주어진 일반적인 특성(지형적 특성, 사회적 특성 등)에 따라 동질적인 부분집합으로 그룹화 될 수 있다. 따라서 각각의 수용자 관측치에 대해 주어진

지형적 특성 내에서 동일지역의 관측치 만이 가능한 제공자로 고려된다. 일반적으로 하나 혹은 몇몇의 범주형 공통변수가 대체군(donation class)이 된다.

바. 평가방법

지금까지의 연구들에서 매칭 결과에 대한 평가들은 각 연구의 특성에 따라 다양하게 제안되었다. 그러나 각각의 방법들을 살펴보면 단순히 분석을 목적으로 표본을 결합할 때 매칭의 성과를 평가하는 방법은 크게 예측력과 대표성의 문제로 압축된다 (van Pelt, 2001).

예측력(정확성)이란 기대되는(또는 알고 있는) 목표와 매칭 결과 사이의 거리 측도로 판단한다. 연속형 변수의 경우 거리의 평균제곱오차(MSE)로, 범주형 변수의 경우 오분류행렬(confusion matrix), 오분류율(error rate) 등을 척도로 하여 정확성을 판단할 수 있다. 또한 대표성이란 매칭결과가 원본 수용파일의 성질을 그대로 유지하는가의 문제를 말한다. 매칭 결과는 원본 표본에서 유일변수와 같은 평균과 표준편차, 그리고 분산을 반드시 가지므로 좋은 간접적인 측도가 된다. 좀 더 직접적인 측도는 결합파일과 수용파일의 결합변수와 고유변수, 그리고 결합파일과 제공파일의 결합변수와 유일변수들의 관계(상관관계, 공분산, 분포)의 비교이다.

2.3 모의실험

모의실험을 위해 다음과 같이 4개의 변수와 정규분포를 가정한다.

$(X_1, X_2, Y, Z) \sim N_4(0, \Sigma)$
여기서

$$\Sigma = \begin{pmatrix} \Sigma_{X_1X_1} & \Sigma_{X_1X_2} & \Sigma_{X_1Y} & \Sigma_{X_1Z} \\ \Sigma_{X_2X_1} & \Sigma_{X_2X_2} & \Sigma_{X_2Y} & \Sigma_{X_2Z} \\ \Sigma_{YX_1} & \Sigma_{YX_2} & \sigma_{YY} & \sigma_{YZ} \\ \Sigma_{ZX_1} & \Sigma_{ZX_2} & \sigma_{ZY} & \sigma_{ZZ} \end{pmatrix} = \begin{pmatrix} 1.0 & 0.2 & 0.5 & 0.8 \\ 0.2 & 1.0 & 0.5 & 0.6 \\ 0.5 & 0.5 & 1.0 & 0.8 \\ 0.8 & 0.6 & 0.8 & 1.0 \end{pmatrix}$$ 이다.

Y와 Z의 공분산은 0.8이며 $X = x$가 주어졌을 때 Y와 Z의 조건부 상관계수는 다음과 같이 구할 수 있다.

$$\rho_{YZ|X} = \frac{\sigma_{YZ|X}}{\sqrt{\sigma_{YY|X}\,\sigma_{ZZ|X}}} = \frac{\sigma_{YZ} - \Sigma_{YX}\Sigma_{ZZ}^{-1}\Sigma_{XZ}}{\sqrt{(\sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})(\sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ})}} = 0.7129$$

여기서 (X_1, X_2, Y)를 수용파일이라 하고, (X_1, X_2, Z)를 제공파일이라 하자. $Y, Z|X = x$의 조건부 독립성이 만족된다면 매칭 후 Y와 Z의 공분산(unconditional covariance)은 다음과 같게 된다.

$$\sigma_{YZ} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ} = (0.5 \quad 0.5)\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix} = 0.5833$$

데이터는 $(X_1, X_2, Y, Z) \sim N_4(0, \Sigma)$로부터 5000개의 난수를 발생시켜 얻는다. 5000개 중 2500개씩 임의로 나누어 각각을 수용파일과 제공파일로 한다. 수용파일에서는 Z를 삭제하고 제공파일에서는 Y를 삭제한다. 매칭방법은 최근접이웃방법을 이용하며, 근사성 측정 시 최소 절대거리를 이용한다. 데이터를 생성시키고, 제공파일과 수용파일로 분할, 두 개의 데이터를 매칭, 그리고 추정치 계산의 전 과정을 50번 반복한다. $E(\widehat{\mu_Z})$, $E(\hat{\sigma}_Z^2)$, $E(\hat{\sigma}_{XZ})$, $E(\hat{\sigma}_{YZ})$, $E(\hat{\rho}_{YZ})$, 그리고 $E(\hat{\rho}_{YZ}|X)$는 다음과 같이 50번의 반복에 대한 단순 평균으로써 추정한다.

$$\hat{E}(\hat{\theta}) = \frac{1}{k}\sum_{j=1}^{k}\hat{\theta}_j$$

추정치 $\hat{\theta}$에 대한 표본표준오차는 다음과 같이 계산한다.

$$s(\hat{\theta}) = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}(\hat{\theta}_j - \hat{E}(\hat{\theta}))^2}, \quad k = 50$$

$\rho_{YZ}|X = 0.7129$인 경우의 모의실험 결과가 <표 2.1>에 나타나 있다. 50번 반복한 결과 $\hat{E}(\hat{\theta})$를 보면 매칭 후의 추정치들이 원래의 데이터의 θ값과 매우 유사한 것을 알 수 있다. 즉, 매칭 후에도 원래의 데이터의 성질을 그대로 유지하고 있다는 것을 알 수 있다. 또한 $\widehat{E}(\hat{\rho}_{YZ|X}) = -0.00486$으로 0에 가까운 값을 가짐으로써 조건부 독립성이 만족되는 것을 확인할 수 있다.

<표 2.1> $\rho_{YZ}|X \neq 0$인 경우 최근접이웃방법을 이용한 매칭결과

반복	$\widehat{\mu_Z}$	$\hat{\sigma}_Z^2$	$\hat{\sigma}_{X_1Z}$	$\hat{\sigma}_{X_2Z}$	$\hat{\sigma}_{YZ}$	$\hat{\rho}_{YZ}$	$\hat{\rho}_{YZ}\|X$
1	0.0131	0.9364	0.7663	0.5522	0.5400	0.5574	-0.0200
2	-0.0086	1.0018	0.8087	0.5927	0.5935	0.5940	0.0045
3	0.0015	0.9880	0.7962	0.5519	0.5754	0.5869	0.0027
4	0.0109	0.9555	0.7403	0.5912	0.5503	0.5712	-0.0032
:	:	:	:	:	:	:	:
47	-0.0141	1.0212	0.8523	0.5975	0.6076	0.5960	0.0381
48	0.0287	0.9613	0.7477	0.5937	0.5636	0.5747	-0.0158
49	-0.0209	1.0025	0.8217	0.6149	0.5807	0.5781	-0.0097
50	-0.0081	1.0352	0.8336	0.6275	0.5802	0.5757	-0.0335
$\hat{E}(\hat{\theta})$	0.0032	0.9888	0.7913	0.5996	0.5804	0.5837	-0.0049
$s(\hat{\theta})$	0.0210	0.0345	0.0311	0.0245	0.0247	0.0131	0.0212

3. 실제자료분석

3.1 사례연구

Boston Housing 데이터(출처: UC Irvine Repository)는 13개의 독립변수를 이용하여 Boston 지역의 집값(MEDV)을 예측하는 것이 목적인 데이터이다.

매칭을 수행하기 위해 원래의 데이터를 수용파일과 제공파일로 각각 분할하여야 한다. 이때 각 파일에 포함될 변수 분리는 조건부 독립성이 만족되도록 이루어져야 한다. 여기서는 Rässler(2002)가 제시한 회귀분석접근법으로 조건부 독립성을 판단하는 방법을 이용한다. 최종분석의 목표변수가 될 MEDV는 데이터 매칭에 영향이 없도록 하기 위해 수용파일의 유일변수 Y에 포함시킨다. 또한 아래와 같이 회귀모형을 가정할 경우 $\beta_{YZ|X}=0$이면 $\rho_{YZ|X}=0$이 성립된다.

$$Z=\beta_0+\beta_{XZ|Y}X+\beta_{YZ|X}Y$$

제공파일의 유일변수 Z가 조건부 독립성이 만족되도록 하기 위하여 MEDV를 설명변수로 유의성이 없는 반응변수 INDUS, AGE, CHAS를 선택(유의수준=0.05)한다. 공통변수 X는 제공파일의 유일변수로 선택된 INDUS, AGE, CHAS변수를 반응변수로 하여 유의한 설명변수 NOX, RM, DIS, RAD, TAX, LSTAT를 선택한다. 수용파일의 유일변수 Y는 공통변수에 포함되지 않고 제공파일의 유일변수에도 포함되지 않는 변수들을 선택한다.

<표 3.1> 실험데이터의 파티션 결과

변수		개체수(506)		공통변수(X)	수용파일 유일변수(Y)	제공파일 유일변수(Z)
연속	범주	수용파일	제공파일			
13	1	202	304	NOX,RM,DIS, RAD,TAX, LSTAT	PTRATIO, MEDV,ZN,B,CRIM	INDUS,AGE, CHAS(범주형)

<표 3.2> 가장 가까운 7개 예측치의 차이 (통합변수가 연속형인 경우)

R	1(INDUS=7.87)			…	100(INDUS=5.86)			…	202(INDUS=27.74)		
k	D	distance	INDUS(D)		D	distance	INDUS(D)		D	distance	INDUS(D)
1	131	0.00316	6.20		165	0.00430	2.25		219	18.10	0.023664
2	137	0.00448	6.20		145	0.01092	4.93		293	27.74	0.139158
3	4	0.00458	7.87	…	148	0.01178	5.86	…	228	18.10	0.185954
4	134	0.01248	6.20		170	0.01627	4.95		294	27.74	0.186472
5	111	0.01576	3.44		146	0.01693	4.93		222	18.10	0.189157
6	195	0.01769	7.38		39	0.01779	5.13		230	18.10	0.237989
7	122	0.02118	10.59		181	0.01781	2.18		85	19.58	0.259588

주: R-수용파일의 개체, D-제공파일의 개체, INDUS(D)-제공파일의 실제값,
distance=$|\hat{St}\_INDUS_R-\hat{St}\_INDUS_D|$

데이터는 수용파일과 제공파일을 각각 60% 대 40%로 하고, 데이터의 분리는 단순임의방법을 사용하였으며, 매칭 알고리즘은 k-근접이웃기법과 회귀분석 기법의 결합기법을 사용하였다. 제공파일에서 연속형 변수 INDUS를 수용파일에 결합해본 결과가 <표 3.2>에 나타나 있다.

표준화한 INDUS를 목표변수로 하고 공통변수를 독립변수로 하여 단계적 변수선택법을 이용한 회귀모형을 수행하였다. 매칭의 수행 결과를 평가하기 위하여 연속형 변수 INDUS에 대해서는 평균제곱오차(MSE)를 사용하였다. 데이터를 분할하고 매칭을 하는 전 과정을 20번 반복하여 시행하였으며, k값에 따른 MSE값의 변화와 20회 반복실험의 MSE값에 대한 평균값이 k가 1에서 7까지 증가하면서 점차 감소하는 것으로 나타났다. 특히 k가 1에서 3으로 증가할 때 MSE의 감소량이 다른 구간에 비해 상당히 큰 것으로 나타났다.

<표 3.3> k에 따른 MSE의 변화 (통합변수가 연속형인 경우)

반복	k=1	k=3	k=5	k=7
1	18.561	15.125	14.292	14.157
2	23.303	15.761	12.221	12.711
3	22.394	13.161	11.122	10.705
4	17.781	15.699	14.492	14.488
5	21.321	15.621	13.386	13.545
:	:	:	:	:
16	18.258	11.363	11.503	10.898
17	24.218	19.710	17.435	14.487
18	20.861	15.841	15.551	15.270
19	24.989	13.937	12.518	10.553
20	20.838	16.960	14.737	13.562
평균	21.523	14.920	13.451	12.941

<표 3.4> k에 따른 오분류율(%)의 변화 (통합변수가 범주형인 경우)

반복	k=1	k=3	k=5	k=7
1	10.8911	3.9604	2.4752	1.4851
2	12.3762	3.9604	1.9802	1.4851
3	8.9109	3.4653	2.4752	1.4851
4	10.8911	2.9703	1.9802	1.4851
5	12.8713	3.9604	2.4752	1.9802
:	:	:	:	:
16	11.3861	3.9604	2.4752	1.9802
17	8.9109	3.4653	1.9802	1.4851
18	12.8713	3.9604	2.4752	1.4851
19	10.8911	3.4653	1.9802	1.4851
20	11.3861	3.9604	2.4752	1.9802
평균	11.5842	3.7624	2.3267	1.6089

제공파일에서 범주형 변수 CHAS를 수용파일에 결합하기 위해 CHAS를 목표변수

로, 공통변수를 독립변수로하여 로지스틱회귀모형을 적합하였다. 데이터 매칭의 수행결과를 평가하기 위해 정확도의 측도로 범주형 변수 CHAS에 대하여 오분류율(error rate)을 사용하였다. 연속형의 경우와 마찬가지로 데이터를 분할하고 매칭을 하는 전 과정을 20번 반복하여 시행하였다. k값에 따른 오분류율의 변화와 20회 반복실험의 오분류율에 대한 평균값이 k가 1에서 7까지 증가하면서 점차 감소하는 것으로 나타났다. 특히 k가 1에서 3으로 증가할 때 오분류율의 감소량이 다른 구간에 비해 상당히 큰 것으로 나타났다.

3.2 통계조사자료와 행정자료 간의 매칭

가. 데이터 설명

국민연금자료는 2006년 6월 기준 서울지역 데이터로 전체 223,186개의 관측치와 18개의 변수를 가지고 있으며, 사업체기초조사자료(이하 사기초자료)는 2005년 12월 기준 서울지역 데이터로 전체 741,229개의 관측치와 72개의 변수로 이루어져 있다.

사기초자료를 수용자 파일로 하고, 국민연금자료를 제공자 파일로 하며, 이때 사기초자료의 '종사자수'를 수용자 파일의 유일변수로 하고, 국민연금자료의 '가입자수'를 제공자 파일의 유일변수로 하여 수용자 파일에 매칭시킨다. 국민연금자료의 가입자수는 평균이 약 13.96이며, 표준편차는 약 267.43, 중위수는 3, 최빈값은 2인 것으로 나타났다. 사기초자료의 종사자수는 평균이 약 5.18이며, 표준편차는 약 36.08, 중위수는 2, 최빈값은 1인 것으로 나타났다.

나. 정확 매칭 I

본 연구에서는 사업자등록번호를 기준변수로 이용하여 정확 매칭을 시행해 본다. 또한 정확 매칭된 데이터로부터 사기초자료의 종사자수와 국민연금자료의 가입자수의 일치율 및 비일치에 따른 데이터의 분포를 파악하고, 사기초자료의 종사자수를 국민연금의 가입자수로 대체하여 사용할 수 있는지 검토한다.

각 데이터를 살펴보면 국민연금자료에는 사업자등록번호가 결측인 관측치가 없었으나, 사기초자료에는 741,229개의 관측치 중 162,681개가 결측인 것으로 나타났다.

<표 3.5> 기준변수의 결측치 제거 후 관측치(1)

구분	국민연금자료	사기초자료
원데이터	223,186개	741,229개
사업자등록번호 결측 제거	223,186개	578,548개
대표자성명 결측 제거	223,171개	578,540개

사업자등록번호를 기준변수로 하여 정확 매칭을 실시한 결과 매칭된 관측치는 296,488개인 것으로 나타났다. 이는 국민연금 원데이터보다 많은 관측치로, 각 데이터에 동일 사업자등록번호를 가지는 관측치들이 다수 존재하여 그들의 가능한 모든 조합으로 매칭이 이루어졌기 때문이다. 또한 데이터를 확인해본 결과 동일 사업자등록번호라 할

지라도 대표자 성명이나 사업체명이 다른 경우가 존재하였다. 따라서 사업자등록번호와 두 데이터 모두에 존재하는 대표자 성명을 기준변수로 추가하여 정확 매칭을 시행하였다. 그 결과 124,826개의 관측치가 정확 매칭되는 것으로 나타났다. 정확 매칭된 124,826개의 관측치를 가지고 국민연금자료의 가입자수와 사기초자료의 종사자수가 일치하는 비율을 살펴본 결과, 종사자수가 증가함에 따라 일치율이 크게 감소하는 경향을 보였다.

정확 매칭된 데이터로부터 종사자수와 가입자수의 관계를 살펴보면 그 차의 값이 매우 큰 경우가 많이 존재하며, 차의 분포가 다양한 것을 알 수 있다. 실제로 규모가 큰 사업체에서는 가입자수와 종사자수의 차도 큰 경향을 보일 가능성이 크므로 각 사업체의 규모 대비 차의 형태로 차 백분율의 분포를 그룹별로 살펴보았다. 그 결과 대부분의 그룹에서 차 백분율이 50% 이상인 경우의 비율이 가장 큰 것으로 나타나 규모 대비 차의 값 또한 큰 것을 알 수 있었다. 이는 조사의 정확성과 응답의 정확성 측면에서 데이터에 대해 검토할 필요가 있다고 판단된다. 또한 각 데이터에는 많은 결측이 존재하여 데이터의 타당성에 대해 검토해 볼 필요가 있는 것으로 판단된다.

다. 정확 매칭 Ⅱ

국민연금자료와 사기초자료 각각에는 법인등록번호가 존재한다. 따라서 사업자등록번호와 대표자 성명을 이용한 정확 매칭I과는 달리 법인등록번호와 대표자 성명을 기준변수로 한 정확 매칭도 고려할 수 있다.

<표 3.6> 기준변수의 결측치 제거 후 관측치(2)

구분	국민연금자료	사기초자료
원데이터	223,186개	741,229개
법인등록번호 결측 제거	115,465개	95,881개
대표자성명 결측 제거	115,462개	95,880개
중복 제거	3,736개	76,575개

<표 3.7> 법인등록번호와 대표자성명을 기준변수로 한 정확 매칭 결과의 예

기준변수		국민연금자료		사기초자료	
법인등록번호	대표자 성명	사업장명칭	업종	사업체명	주사업내용
11011*****	이**	롯데쇼핑(주)	소매업	롯데 디자인팀	인테리어 디자인업
11011*****	이**	롯데쇼핑(주) 롯데시네마	오락,문화,운동관련	롯데 디자인팀	인테리어 디자인업
11011*****	이**	롯데쇼핑(주) KKD사업본부	숙박,음식업	롯데 디자인팀	인테리어 디자인업

국민연금자료에서 법인등록번호와 대표자성명이 결측인 관측치를 제외하면 115,462개의 관측치를 얻을 수 있으며, 사기초자료에서 법인등록번호와 대표자성명이 결측인 관측치를 제외하면 95,880개의 관측치를 얻을 수 있다. 이 두 자료를 법인등록번호와

대표자 성명을 기준변수로 하여 정확 매칭을 시행하면 서로 다른 업체가 매칭이 되는 문제가 발생한다. 동일 사업자등록번호와 동일 대표자 성명을 가진 경우지만 다른 업체가 매칭되는 것을 알 수 있다. 이는 다수의 업체가 동일 법인등록번호와 대표자 성명을 가지고 있기 때문이다. 실제로 각 데이터에서 중복 법인등록번호를 제거하면 국민연금자료의 경우 3,736개의 관측치만이 존재하고, 사기초자료의 경우 76,575개의 관측치만이 존재하는 것으로 나타났다. 따라서 법인등록번호와 대표자성명을 이용한 정확 매칭은 타당하지 않다고 할 수 있다.

라. 통계적 매칭

사기초자료를 수용자 파일로 하고, 국민연금자료를 제공자 파일로 하며, 이때 사기초자료의 '종사자수'를 수용자 파일의 유일변수로 하고, 국민연금자료의 '가입자수'를 제공자 파일의 유일변수로 하여 수용자 파일에 매칭시킨다. 그런데 주어진 사업체기조사자료와 국민연금자료간의 통계적 매칭을 시행하기에는 두 데이터 간의 공통변수가 부족하다. 국민연금자료와 사기초자료를 보면 정확 매칭이 되는 대표자성명, 사업자등록번호를 제외하고 다음의 변수들을 공통변수로 예상해 볼 수 있으나 각각의 문제점이 있다.

소재지: 사기초자료에는 소재지가 사업체_읍면동, 소재지_사업체주소지 등의 변수로 구분되어 있으나, 국민연금자료에는 소재지가 하나의 텍스트 문장으로 되어 있어 사기초자료에서와 같이 구분하기가 힘들다.

사업장형태: 국민연금자료에는 '법인', '개인'의 값을 가지나 사기초자료에는 조직형태라는 변수명으로 '개인사업체', '회사법인', '회사 외 법인', '국가·지방자치단체', '비법인 단체'의 값을 가진다. 따라서 두 데이터의 속성을 동일하게 하여 하나의 공통변수로 만들기 위해서는 사업장형태 변수의 범주에 대한 통일이 필요하다.

업종: 국민연금자료에는 63가지의 업종으로 분류되어 코딩되어 있으나(결측 52개), 사기초자료에는 '사업의 종류_주사업내용' 이라는 변수가 있는데 너무 많은 수의 범주(종류)가 있어 구분하기가 힘들다.

또한 각 데이터의 관측치 개수를 살펴보면 수용자 파일로 사용될 사기초자료가 제공자 파일로 사용될 국민연금자료보다 훨씬 크다는 문제점이 있다. 따라서 정확 매칭된 데이터 124,826개를 가지고 다시 제공자 파일과 수용자 파일로 나누어 통계적 매칭을 시행해 본다. 이는 후에 매칭에 대한 평가가 용이하다는 장점이 있다. 또한 공통변수의 부족 문제도 해결된다. 소재지 변수는 사기초자료에 있는 변수를 쓰고, 업종은 국민연금자료의 변수를 써서, 다시 이들을 둘로 나누면 소재지와 업종이라는 공통변수를 사용할 수 있을 것이다. 실제로 많은 시간을 들여 국민연금의 소재지변수를 가공하고, 사기초자료의 '사업의 종류_주사업내용'을 가공하면 동일한 결과를 얻을 것으로 기대된다. 또한 국민연금자료의 사업장 형태와 사기초자료의 조직형태에 대한 범주 통일을

위해 분할표를 작성해본 결과, 분할표의 칼럼백분율을 바탕으로 조직형태의 값이 '개인사업체'인 경우는 개인으로 볼 수 있고, '회사법인', '회사 외 법인', '국가·지방자치단체'의 경우는 법인으로 볼 수 있다. 조직형태의 값이 '비법인 단체'인 경우는 구분이 명확하지 않아 이들 관측치 1,880개는 통계적 매칭에서 제외하기로 한다.

따라서 세 개의 범주형 공통변수(소재지, 업종, 사업형태)를 이용하여 통계적 매칭을 할 수 있다. 그러나 공통변수로 사용할 수 있는 변수의 개수가 적고 모두 범주형인 경우 데이터 매칭 시 많은 동점이 방생할 가능성이 크므로, 데이터 매칭 방법 중 활용가치가 큰 랜덤 핫덱 방법을 이용하여 수용자 파일의 관측치들이 동일한 값을 갖더라도 제공자 파일에서 상이한 관측치들이 매칭되게 하여 변동이 발생되도록 한다.

사업자등록번호와 대표자성명을 기준으로 하여 정확 매칭한 데이터 124,826개에서 조직형태가 명확하지 않은 관측치 1,880개를 제거하고, 후의 평가를 위해 공통변수(주소, 조직형태, 사업내용)가 불일치하는 관측치 2,420개를 제거한 120,526개의 데이터에 대해 랜덤 핫덱 방법을 적용하였다.

<표 3.8> 사업장 형태와 조직형태의 분할표

사업장 형태	조직형태					총합
	개인 사업체	회사법인	회사 외 법인	국가·지방 자치단체	비법인 단체	
개인	58241 (96.40%)*	89 (0.16%)	122 (2.79%)	34 (1.36%)	743 (39.52%)	59229
법인	2175 (3.60%)	55560 (99.84%)	4257 (97.21%)	2486 (98.64%)	1137 (60.48%)	65597
총합	60416	55649	4379	2502	1880	124826

주: ( )안 %는 칼럼 백분율임.

<표 3.9> 실제값과 매칭값의 차이값에 대한 분포

차이값	빈도	백분율	누적 빈도	누적 백분율
0	38497	31.94	38497	31.94
1	19785	16.42	58282	48.36
2	13163	10.92	71445	59.28
3	8669	7.19	80114	66.47
4	5944	4.93	86058	71.40
5	4298	3.57	90356	74.97
6-10	11104	9.21	101460	84.18
11-20	7663	6.36	109123	90.54
21-30	2924	2.43	112047	92.97
31-40	1586	1.32	113633	94.28
41-50	998	0.83	114631	95.11
51-60	720	0.60	115351	95.71
61-70	500	0.41	115851	96.12
71-80	402	0.33	116253	96.45
81-90	320	0.27	116573	96.72
91-100	265	0.22	116838	96.94
101이상	3688	3.06	120526	100.00

통계적 매칭 결과는 앞에서 언급한 바와 같이 대표성 측면과 정확성 측면에서 평가해볼 수 있다. 원래의 국민연금자료에서의 가입자수의 분포와 통계적 매칭이 된 데이터에서의 가입자수의 분포를 비교해 두 분포가 유사하면 매칭결과가 원본 파일의 성질을 잘 유지하고 있다고 할 수 있다. 실제로 제공자 파일에서의 가입자수의 분포와 매칭된 파일에서의 가입자수의 분포를 비교해본 결과, 평균은 각각 57.19와 59.28로 큰 차이가 나지 않으며, 중위수는 동일한 것으로 나타났다. 표준편차와 MAE 역시 큰 차이가 나지 않는 것으로 나타났다.

따라서 매칭된 파일에서의 가입자수의 분포가 제공자 파일의 가입자수의 분포를 그대로 유지하고 있어 대표성 측면에서 매칭 결과가 타당하다고 볼 수 있다. 또한 정확 매칭된 데이터를 바탕으로 매칭의 결과가 정확한지 평가해보면 실제값과 매칭에 의한 값이 정확하게 일치하는 경우는 31.94%에 불과하나 차이가 5 이하인 경우는 전체의 74.97%로 비교적 높은 것을 알 수 있다. 따라서 정확성 측면에서도 매칭결과가 타당하다고 판단된다. 소수의 범주형 공통변수를 사용하여 랜덤 핫덱 방법을 적용하였음에도 대표성과 정확성 측면에서 신뢰할 만한 결과를 얻었다고 볼 수 있다.

[수용자파일-사기초자료]

공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (종사자수)
신정3	교육 서비스업	법인	79
신정3	교육 서비스업	법인	85
신정3	교육 서비스업	법인	85
신정3	교육 서비스업	법인	7
신정3	교육 서비스업	법인	70

+

[제공자파일-국민연금조사자료]

공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (종사자수)
신정3	교육 서비스업	법인	18
신정3	교육 서비스업	법인	6
신정3	교육 서비스업	법인	5
신정3	교육 서비스업	법인	16
신정3	교육 서비스업	법인	16

=

[매칭된 파일]

공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (종사자수)	매칭된 변수 (가입자수) : A	정확매칭데이터 (가입자수): B	차이 |A - B|
신정3	교육서비스업	법인	79	16	18	2
신정3	교육서비스업	법인	85	5	6	1
신정3	교육서비스업	법인	85	6	5	1
신정3	교육서비스업	법인	7	16	16	0
신정3	교육서비스업	법인	70	16	16	0

<그림 3.1> 랜덤 핫덱 방법 적용 결과 (예시)

4. 결론 및 향후 연구과제

국민연금자료와 사기초자료에 대해 사업자등록번호와 대표자 성명을 기준변수로 사용하여 정확 매칭을 적용시켜보았다. 정확 매칭된 데이터로부터 사기초자료의 종사자수와 국민연금자료의 가입자수의 일치율 및 비일치에 따른 데이터의 분포를 파악해 본 결과 종사자 그룹에 따른 일치율이 크게 차이가 있으며, 종사자수가 증가함에 따라 일치율이 크게 감소하는 것으로 나타났다. 또한 가입자수와 종사자수의 차의 값이 매우 다양하게 분포하고 있음을 확인할 수 있었다.

통계적 매칭의 경우 사기초자료를 수용자 파일로 하고, 국민연금자료를 제공자 파일로 하였으며, 사기초자료의 '종사자수'를 수용자 파일의 유일변수로, 국민연금자료의 '가입자수'를 제공자 파일의 유일변수로 하여 수용자 파일에 매칭시켰다. 이때 두 데이터에 모두 존재하는 소재지, 업종, 조직형태의 3가지 범주형 변수를 공통변수로 사용하였다. 제공자 파일인 국민연금자료에서의 가입자수의 분포와 통계적 매칭이 된 데이터에서의 가입자수의 분포를 평균, 중위수, 표준편차, MAE등의 기준으로 비교해 본 결과 두 분포가 거의 유사한 것을 알 수 있었다. 따라서 통계적 매칭결과가 원본 파일의 성질을 잘 유지하고 있다고 할 수 있다. 또한 정확 매칭된 데이터를 바탕으로 통계적 매칭의 결과에 대한 정확성을 평가해 본 결과 실제값과 매칭에 의한 값이 정확하게 일치하는 경우는 31.94%에 불과하나 차이가 5 이하인 경우는 전체의 74.97%로 비교적 높은 것을 알 수 있었다. 따라서 통계적 매칭의 경우 소수의 공통변수(소재지, 업종, 조직형태)를 사용하여 랜덤 핫덱 방법을 적용하였음에도 대표성과 정확성 측면에서 신뢰할 만한 결과를 얻었다고 할 수 있다.

본 연구에서는 데이터에 많은 결측치가 존재하며, 데이터 매칭에 있어 공통변수로 사용가능한 변수의 수가 적다는 것이 문제점으로 지적되었다. 따라서 향후 국민연금자료 및 사기초자료에 대한 추가적인 연구가 필요한 것으로 판단된다. 국민연금자료에서 '가입대상자수(가입자수)'외에 '근로자수'를 활용하는 방안과 국민연금의 '근로자수'와 사업체기초조사의 '종사자수'의 개념 및 데이터에 대한 비교가 필요하다. 또한 국민연금 사업장 폐쇄신고 등의 변동데이터에 대해서도 활용 방안을 연구해 볼 필요가 있을 것이다. 정확 매칭을 위한 변수 간 조건을 개발하고, 통계적 매칭을 위한 국민연금자료의 추가 변수를 활용하는 것도 데이터 매칭의 효율성을 높일 수 있을 것이라 기대된다.

감사의 글

이 연구는 2007년 통계개발원의 연구용역과제로 수행되었음.

참고문헌

D'Orazio, Marcello, Di Zio, Marco and Scanu, Mauro. (2006). Statistical Matching Theory and Practice, Wiley.

Rässler, S. (2002). Statistical Matching: A frequentist theory, practical applications, and alternative Bayesian approaches, Springer Verlag, New York.

_____ (2004). Data fusion : identification problem, validity, and multiple imputation, *Austrian Journal of Statistics*, 33, 153-171.

Van Der Putten, Peter, Kok, Joost N., and Gupta, Amar. (2002). Why the information explosion can be bad for data mining, and how data fusion provides a way out, Second SIAM International Conference on Data Mining, Arlington, April, 11-13.

_____ (2002). Data Fusion through Statistical Matching, Technical Paper, 185, Center for eBusiness@MIT, MITSloan (ebusiness.mit.edu)

van Pelt, X. (2001). The Fusion Factory : A Constrained Data Fusion Approach. Master of Science, Thesis, Leiden Institute of Advanced Computer Science, The Netherlands.

National Statistics (2003). National Statistics code of Practice-Protocol on Data Matching, London : TSO.

(2009년 2월 17일 접수, 2009년 3월 27일 수정, 2009년 4월 2일 채택)

A Study on the Statistical Matching between Survey Data and Administrative Data

Yung-Seop Lee[1] · Sun-Woong Kim[2] · Hong-Yup Ahn[3] · Kyung-Eun Lim[4] · Hee-Kyung Kim[5]

Abstract

Data fusion by the statistical matching technique is a way to obtain information from two or more data sources when a data does not contain necessary variables or contain missing values. When we analyze the data using statistical techniques, it is hard to have a data set which have sufficient variables. A variable matched data set from the several data sources is necessary to remedy this situation using various statistical methods. It can improve the data quality, and also reduce the cost and time of data gathering. The objective of statistical matching techniques is the generation of a new data set allows even more flexible analysis than each single data set. In our study, we discuss statistical matching techniques and assessment methods by statistically matching a survey data and a administrative data. In case study, both 'National pension data' and 'The census on basic characteristics of establishments' data are statistically matched using the random hot deck method. It is also validated by comparing with the original data. In the future, we can apply the statistically-matched data set for more efficient data analysis and practical usage.

Key words : data fusion, statistical matching, recipient file, donor file, random hot deck

1) Associate Professor, Dongguk University. E-mail: yung@dongguk.edu
2) Associate Professor, Dongguk University. E-mail: sunwk@dongguk.edu
3) Assistant Professor, Dongguk University. E-mail: ahn@dongguk.edu
4) (Corresponding author) Deputy Director, Statistics Research Institute. E-mail: kelim06@nso.go.kr
5) Ph.D. Candidate, Dongguk University. E-mail: khk0228@dongguk.edu