Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 4, pp. 879 ~ 890

A Statistical Matching Method with k-NN and Regression[1)]

Sung S. Chung[2)] · Soon Y. Kim[3)]
Seung S. Lee[4)] · Ki H. Lee[5)]

Abstract

Statistical matching is a method of data integration for data sources that do not share the same units. It could produce rapidly lots of new information at low cost and decrease the response burden affecting the quality of data. This paper proposes a statistical matching technique combining k-NN (k-nearest neighborhood) and regression methods. We select k records in a donor file that have similarity in value with a specific observation of the common variable in a recipient file and estimate an imputation value for the recipient file, using regression modeling in the donor file. An empirical comparison study is conducted to show the properties of the proposed method.

Keywords : Data Fusion, k-NN, Statistical Matching

1. 서론

통계적 매칭(데이터 통합)은 보유하고 있는 데이터 파일에 필요한 변수가 없거나, 결측값이 존재할 경우 다른 원천으로부터 모아지는 데이터와 정보(information)를 통합시키는 것으로 정의된다. 통계적 매칭은, 기존 자료로부터 정보를 얻을 때 우리가

1) 이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 일부 지원을 받아 수행된 연구임(No. R01-2005-000-10752-0).

2) 전북 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보학과(응용통계연구소) 교수
E-mail: sschung@chonbuk.ac.kr

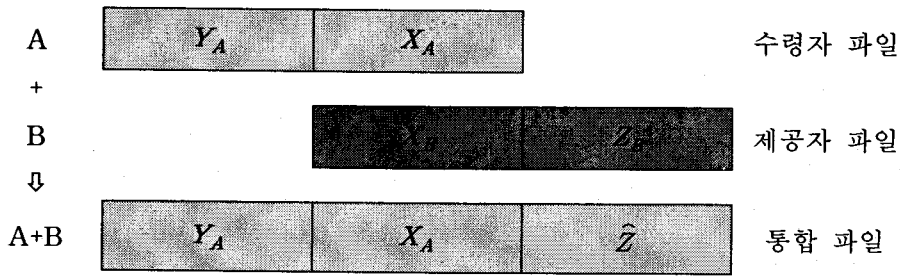
3) 전북 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보학과 박사과정
E-mail: rabbit@chonbuk.ac.kr

4) 전북 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보학과 박사과정
E-mail: pinokio129@hanmail.net

5) 교신저자 : 전북 전주시 완산구 효자동 3가 1200, 전주대학교 경영학부 교수
E-mail: khlee@jj.ac.kr

필요로 하는 변수를 모두 포함하는 데이터 파일이 흔하지 않기 때문에, 필요한 변수를 포함하고 있는 여러 개의 파일을 합하여 데이터를 보강하는 수단으로 사용되어 왔다. 이는 필요한 데이터를 재조사하는 방법보다 비용과 시간을 절약할 수 있을 뿐만 아니라, 때로는 더욱 신뢰성이 높은 방법이 될 수 있으며, 조사 응답자의 부담을 줄여줌과 동시에 응답자로부터 성의 있는 응답을 기대할 수 있는 기법이다. 데이터 통합은 1960년대에 유럽과 호주의 미디어 리서치(media research) 분야에서 매우 인기 있고 쟁점이 되는 주제였고 1970년대에는 미국과 캐나다에서 상업적인 분야 외에도 공적인 센서스에 이용되기 시작하였다. 현재까지도 데이터 통합은 여러 종류의 통계적 조사(survey)에서 응답자 수와 설문 문항 수를 줄이는데 꾸준히 이용되고 있다(Rässler(2002)).

데이터 통합의 구조를 살펴보면 다음 <그림 1>과 같다.



<그림 1> 데이터 통합 개념도

통합과정에 의해 확장될 데이터 파일을 수령자 파일(recipient file), 통합을 위해 사용될 추가 정보를 갖는 데이터 파일을 제공자 파일(donor file)이라 한다. 두 데이터 파일에 공통적으로 포함하는 변수를 공통변수(common variables)라 하며 X로 표시한다. 그리고 각 파일에 유일하게 존재하는 변수를 유일변수(unique variables)라 하며 수령자 파일에서는 Y, 제공자 파일에서는 Z로 표시한다. 데이터 통합은 제공자 파일의 정보를 이용하여 제공자 파일의 유일변수 Z를 수령자 파일에 추가하여 통합된 파일(fused file)을 완성하는 작업을 의미한다. 이 통합과정을 통해 수령자 파일에 추가된 변수들을 통합변수(fusion variables)라 하며 $\hat{Z}$로 표시한다.

데이터 통합의 접근법은 통합파일의 분포함수와 그 모수를 추정하는 매크로(macro) 접근법과 수령자 파일에 존재하지 않은 유일변수 Z를 추정하여 추정된 Z를 수령자 파일에 추가하여 완성된 통합파일을 얻는 마이크로(micro) 접근법으로 분류된다. 이 두 접근법은 서로 배타적인 것이 아니고 일반적으로 매크로 방법에 의하여 얻은 정보를 이용하여 마이크로 방법을 적용하는 순차적 단계로 볼 수 있다. 본 논문에서는 마이크로 접근법에 의하여 통합파일을 얻고 이러한 통합결과가 역으로 매크로 접근법에서 필요로 하는 분포모수를 얼마나 정확히 추정하는가에 관심을 갖고자 한다.

본 논문에서 고려할 마이크로 데이터 통합기법은 수령자 파일의 한 개체와 제공자 파일의 모든 개체사이의 유사성을 계산하여 가장 유사한 제공자 파일의 k 개체를 선택하고 이를 이용하여 수령자 파일에 정보를 추가하는 k-최근접이웃법(k-NN:

k-nearest neighborhood)과 제공자 파일에서 X와 Z를 이용하여 회귀식을 구한 후 이 회귀식을 수령자 파일에 적용시켜 조건부 평균(conditional mean)값을 대체하는 회귀분석법(regression) 등의 두 가지 방법과 그들의 결합된 방법 등이다. 2장에서는 회귀분석과 k-NN 방법의 알고리즘과 특징을 살펴보고 이 두 방법을 결합한 기존의 방법과 새롭게 제안하는 결합법을 소개할 것이다. 3장에서는 제안한 방법과 기존의 방법의 특성을 모의실험을 통하여 비교하고, 4장에서 결론과 제언을 한다.

2. 데이터 통합 알고리즘

2.1 회귀분석 알고리즘

회귀분석을 이용하는 방법은 제공자 파일에서 회귀모형을 추정한 후, 추정된 회귀모형을 이용하여 수령자 파일과 제공자 파일에서 예측치를 구한다. 그리고 두 파일의 예측치 사이의 거리가 가장 가까운 개체를 이용하여 통합을 수행하는 방법이다. 이는 Kadane(1978)과 Rubin(1986) 등에 의하여 통계적 매칭에 처음 사용되었고, Singh et al.(1993), Moriarity와 Scheuren(2001, 2003) 등에 의해 일반화와 확장이 행하여졌고, D'Orazio et al.(2006)에 의해 정리되었다.

우리가 추정하여야 할 모수는 다음과 같다.

$$\theta = (\mu, \Sigma) = \left[\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix}, \begin{pmatrix}\Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ}\\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ}\\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ}\end{pmatrix}\right].$$

회귀분석법의 첫 번째 단계로 제공자 파일의 유일변수 Z를 목표변수로, 제공자 파일의 공통변수 X를 설명변수로 하여 회귀모형을 추정한다.

제공자파일에서 X에 대한 Z의 회귀방정식은 식(2.1)과 같다.

$$Z_B = \alpha_Z + \beta_{ZX} X_B + e_{Z|X}, \tag{2.1}$$

여기서 $\alpha_Z = \mu_Z - \beta_{ZX}\mu_X$, $\beta_{ZX} = \Sigma_{ZX}\Sigma_{XX}^{-1}$ 이고, $e_{Z|X} \sim MN(0, \Sigma_{ZZ|X})$와 같이 정규분포를 가정한다. 단, $\Sigma_{ZZ|X} = \Sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ}$이다 .

다음 단계로 제공자 파일에서 추정된 회귀모형을 수령자 파일과 제공자 파일에 각각 적용시켜 유일변수 Z의 예측값 $\tilde{Z}$를 구한다.

$$\begin{aligned}\tilde{Z}_a^A &= \hat{\alpha}_Z + \hat{\beta}_{ZX} X_a^A, \quad a = 1, \cdots, n_A\\ \tilde{Z}_b^B &= \hat{\alpha}_Z + \hat{\beta}_{ZX} X_b^B, \quad b = 1, \cdots, n_B\end{aligned} \tag{2.2}$$

여기서 $\hat{\alpha}_Z$와 $\hat{\beta}_{ZX}$는 파일 B에서 구한 α_Z와 β_{ZX}의 최대우도(ML)추정량이다.

이 예측값 $\tilde{Z}$을 이용하여 수령자 파일의 각 개체와 제공자 파일의 모든 개체사이의

거리를 계산한다. 예측값 사이의 거리가 가장 가까운 제공자 파일의 하나의 개체의 유일변수 관측값 Z를 수령자 파일의 해당 개체에 추가한다. 즉,

$$\text{Min}_b | \tilde{Z}_a^A - \tilde{Z}_b^B | = | \tilde{Z}_a^A - \tilde{Z}_{b^*}^B | \tag{2.3}$$

일 때 다음과 같이 대체한다.

$$\hat{Z}_a^A \leftarrow Z_{b^*}^B.$$

이보다 일반적인 방법으로는 수령자파일의 공통변수 X와 유일변수 Y 를 이용하여 추정된 회귀식을 제공자 파일에 적용시켜 제공자 파일에서도 Y를 추정한다. 그리하여 수령자 파일의 $(\tilde{Z}, Y)$와 제공자 파일의 $(Z, \tilde{Y})$의 거리를 식 (2.3)과 같은 방식으로 계산하여 수령자 파일의 각 개체와 가장 유사한 제공자 파일의 개체를 선택하여 통합에 이용한다. Moriarity와 Scheuren(2001)은 수령자 파일에서 식 (2.2)와 같이 $\tilde{Z}^A$를 추정하여 각 추정값 $\tilde{Z}^A$와 Z^B의 거리가 가장 짧은 개체를 선택해서 그때 $Z_{b^*}^B$값으로 대체하여 주는 방법을 제안하였다.

다음에서 설명할 최근접이웃기법은 데이터 통합과정에서 공통변수 X의 정보만을 이용하나, 회귀분석 기법은 공통변수 X와 제공자 파일의 유일변수 Z의 정보를 이용하므로 데이터 통합방법에 회귀분석과 같은 예측평균매칭(predicted mean matching) 기법이 좋은 성능을 나타낸다(Ingram et al.(2000)). 그러나 모든 개체를 대상으로 회귀식을 얻기 때문에 실제 현실 자료에서는 결정계수 R^2이 매우 적은 것이 일반적이고 이러한 적합도가 낮은 회귀식을 이용해 예측치를 구하면 공통변수 X와 회귀식이 연관성이 적어져 좋은 추정값을 구하기 어렵다. 이는 3장의 모의실험결과에서도 확인할 수 있다.

2.2 k-최근접이웃 알고리즘

k-최근접이웃방법(k-nearest neighborhood method)은 수령자 파일의 한 개체에 가장 가까운 제공자 파일의 k개의 개체를 선택하고 이들의 정보를 이용하여 통합하는 방법이다. 이는 Eubank(1988)의 아이디어를 Härdle(1992)이 커널함수를 이용하여 일반화한 방법으로 비모수적인 방법의 특정한 경우라 볼 수 있다. 이 추정법의 특성으로는 Cheng과 Chu(1996)가 커널추정값의 일치성을 증명하였고, Nielson (2001)이 비모수적 마이크로 접근법으로서의 유용성을 보였다.

알고리즘은 공통변수 X를 이용하여 수령자 파일의 각 개체에 대해 제공자 파일의 모든 개체와의 거리를 계산하는 단계부터 시작한다.

$X = X_a^A$가 주어졌을 때 X_b^B중에서 X_a^A와 가장 가까운 제공자 파일의 k개의 개체를 선택한 후, 선택된 k개 개체에 해당하는 제공자 파일의 유일변수 Z를 이용하여 수령자 파일의 각 개체에 통합 변수를 추가시킨다. 이때 유일변수가 연속형인 경우 k개 Z값의 평균(mean)을, 범주형이면 k개 Z값의 최빈값(mode)을 이용한다. 유일변수가 연속형인 경우 다음과 같이 표현할 수 있다.

$$\hat{Z}_a^A = \frac{1}{k}\sum_{b=1}^{N_B} W_{kb}(\boldsymbol{x})\, Z_b \ , \quad W_{kb}(\boldsymbol{x}) = \begin{cases} 1, & b \in J_{\boldsymbol{x}} \\ 0, & \text{기타} \end{cases}$$

여기서, $J_{\boldsymbol{x}}$는 $\boldsymbol{X}_a^A$와 가장 유사한 제공자 파일에서 선택된 k개 개체의 집합이다.

2.3 회귀분석과 k-NN 알고리즘

앞서 설명한 회귀분석법과 k-NN방법을 결합한 알고리즘은 정성석 외(2004)가 제안한 방법이다. 이를 편의상 회귀분석후 k-NN 방법이라 부르겠다. 이는 회귀분석 알고리즘에 의해 예측값을 구한 뒤 이들 예측값과 가까운 k개의 제공자 파일의 $\boldsymbol{Z}$를 선택하여 이들의 평균 등으로 대체하는 방법이다.

단계별로 보면 처음으로 회귀분석법 알고리즘과 동일하게 제공자 파일의 유일변수 $\boldsymbol{Z}$를 목표변수로, 제공자 파일의 공통변수 $\boldsymbol{X}$를 설명변수로 하여 식 (2.2)와 같이 회귀모형을 추정하고 추정된 회귀모형을 수령자 파일과 제공자 파일에 각각 적용하여 유일변수 $\boldsymbol{Z}$의 예측값 $\tilde{\boldsymbol{Z}}$를 구한다.

$\tilde{Z}_b^B$ 중에서 $\tilde{Z}_a^A$와 가장 가까운 제공자 파일의 k개의 개체를 선택한 후, 선택된 k 개 개체에 해당하는 제공자 파일의 유일변수 $\boldsymbol{Z}$를 이용하여 수령자 파일의 각 개체에 통합 변수를 추가시킨다. 이때 유일변수가 연속형인 경우 k개 $\boldsymbol{Z}$값의 평균(mean)을, 범주형이면 k개 $\boldsymbol{Z}$값의 최빈값(mode)을 이용한다. 유일변수가 연속형인 경우 다음과 같이 표현할 수 있다.

$$\widehat{Z_a^A} = \frac{1}{k}\sum_{b=1}^{n_B} W_b(z)\, Z_b \ , \quad W_b(z) = \begin{cases} 1\ , & b \in J_z \\ 0\ , & otherwise \end{cases} .$$

여기서, J_z는 $\tilde{Z}_a^A$와 가장 유사한 제공자 파일에서 선택된 k개 개체의 집합이다.

이 방법은 k=1일 때 회귀분석 알고리즘과 동일하다는 점에서 알 수 있듯이 k-NN 방법과 달리 공통변수 $\boldsymbol{X}$의 정보를 충분히 이용하지 못한다는 단점이 있다. 그러나 뒤의 모의실험결과에서 알 수 있듯이 회귀분석방법, k-NN 방법보다는 대체로 우수하다고 할 수 있다.

2.4 제안하는 h-NN과 회귀분석 알고리즘

앞 절에서 회귀분석법과 k-NN방법을 결합한 알고리즘은 회귀분석에 의하여 예측치를 구한 뒤 이와 유사한 개체를 k-NN으로 선택하는 방법이었다. 그러나, 이 방법 또한 실제 자료에서 결정계수가 매우 낮게 나올 때는 공통변수 $\boldsymbol{X}$의 정보를 충분히 이용하지 못하는 단점이 있을 수 있다. 본 논문에서 제안하는 h-NN과 회귀분석(regression on neighborhood) 알고리즘은 공통변수 $\boldsymbol{X}$의 유사성을 바탕으로 h-NN을 이용하여 h개의 개체를 선택하여 이들 축소된 개체들을 이용하여 회귀식을 얻은 뒤 이 추정된 회귀식을 수령자 파일과 제공자 파일에 적용하여 예측값을 추정하는 형식이다. 수령자 각 개체별로 h개의 제공자 개체를 구하고 이들의 회귀식을 얻기 때문에

모두 n_A번의 회귀식 계산이 필요하여 컴퓨팅 능력이 매우 필요하다. 하지만 컴퓨터의 발달로 회귀식을 과거에 비해 훨씬 수월하게 얻을 수 있다.

제안한 알고리즘의 첫 번째 단계는 $X = X_a^A$가 주어졌을 때 X_b^B중에서 X_a^A와 가장 가까운 제공자 파일의 h개의 개체 $X_b^B (b \in J_{x_a})$를 선택한다. 그리고 J_{x_a}에서 얻은 회귀식을 이용하여 수령자 파일 A의 a번째 개체의 유일변수 Z의 예측값 $\tilde{Z}$를 구한다.

$$\tilde{Z}_a^A = \hat{\alpha}_Z + \hat{\beta}_{ZX} X_a^A \ .$$

여기서,$\hat{\alpha}_Z$와 $\hat{\beta}_{ZX}$는 X_a^A와 가장 가까운 제공자 파일의 h개의 개체 집합 J_{x_a}을 이용하여 구한 α_Z와 β_{ZX}의 최대우도추정량이다.

마지막 단계로 예측값 사이의 거리가 가장 가까운 제공자 파일의 하나의 개체의 유일변수 관측값 Z를 수령자 파일의 해당 개체에 추가한다. 즉,

$$\mathrm{Min}_b | \tilde{Z}_a^A - \tilde{Z}_b^B | = | \tilde{Z}_a^A - \tilde{Z}_{b^*}^B |$$

일 때 다음과 같이 대체한다.

$$\hat{Z}_a^A \leftarrow Z_{b^*}^B .$$

이 방법은 h를 결정함에 있어 공통변수의 차원과 제공자 파일의 개체수를 동시에 고려한다. 설정된 모의실험의 h수는 제공자 파일의 개체수가 202개이므로 제공자 파일의 10%와 공통변수의 수가 6개임을 동시에 고려해 최소 20개의 관측치가 있어야 할 것으로 판단하였다. 이처럼 이 방법은 공통변수 X의 차원에 따라 h값이 매우 커지는 것이 단점이라 할 수 있다. 이에 대한 해결책은 4장의 결론 부분에서 언급한다.

3. 모의실험

모의실험을 통해 앞서 소개한 네 가지 알고리즘의 특성을 비교하고자 한다. 일반적인 모의실험 비교를 위해 정규난수를 생성하고 각 방법을 적용하여 그 예측의 정확도를 추정해볼 수 있지만 D'Orazio et al.(2006)의 모의실험 결과에 의하면 이론적인 가정에 잘 맞는 경우 각 알고리즘의 정확성이 거의 비슷하게 나와 그 특성을 살피는데 한계가 있다. 그래서 본 논문에서는 통계학교재에서 즐겨 인용되는 Boston Housing 자료를 이용하여 모의실험을 실행해 보았다. 보스턴 주택자료는 보스턴 지역의 주택 관련 14개 변수(13개 계량변수, 1개의 이항변수)에 관하여 506개의 관측값을 얻은 것이다. 이를 Yoshizoe(1996)가 제안한 대로 수령자와 제공자의 비율을 6:4로 나누고 200번 반복하여 통합된 파일의 유일변수 Y와 통합된 변수 $\hat{Z}$의 공분산을 추정하고 수령자 파일의 유일변수Y와 통합될 변수의 실제값 Z의 공분산을 구해 이 두값의 차

이인 MSE를 구하였다.

우선 공통변수와 유일변수를 선택함에 있어 다음과 같은 조건부독립가정(CIA: conditional independence assumption)이 최대한 만족되도록 하였다.

$$\Sigma_{YZ|X} = 0 .$$

조건부 독립가정은 매크로 접근법에서 동시에 측정된 적이 없는 변수들의 공분산, 즉 여기서는 Σ_{YZ}을 추정할 때 필요한 가정이다. CIA 가정이 만족되어야만 다음과 같이 Σ_{YZ}을 추정할 수 있기 때문이다.

$$\hat{\Sigma}_{YZ} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ} .$$

본 논문에서는 마이크로 접근법에 의해 미관측된 변수값을 대체해주지만 최종적으로 관심이 있는 것은 대체값에 의하여 Σ_{YZ}를 얼마나 정확히 추정하는가 이기 때문에 이러한 가정이 가능한 만족하도록 실험을 설계하였다. 그리고 CIA 가정이 필요한 다른 이유로 다음과 같이 회귀분석에 의한 대체법인 경우 조건부평균값에 오차항을 더해줌으로써 Σ_{YZ}를 더 정확하게 추정해줄 수 있다.

$$\hat{Y} = \mu_Y + \Sigma_{XY}\Sigma_{XX}^{-1}(X-\mu_X) + e_{Y|X} .$$
$$\hat{Z} = \mu_Z + \Sigma_{XZ}\Sigma_{XX}^{-1}(X-\mu_X) + e_{Z|X} .$$

여기서, $e_{Y|X} \sim MN(0, \Sigma_{Y|X})$, $e_{Z|X} \sim MN(0, \Sigma_{Z|X})$, $\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{XY}\Sigma_{XX}^{-1}$, $\Sigma_{Z|X} = \Sigma_{ZZ} - \Sigma_{XZ}\Sigma_{XX}^{-1}$ 이다.

이 경우 Y, Z의 공분산을 구해보면 다음과 같다.

$$\begin{aligned} Cov(Y,Z) &= E(Y-E(Y))(Z-E(Z))' \\ &= E(\Sigma_{XY}\Sigma_{XX}^{-1}(X-\mu_X)+e_{Y|X})((X-\mu_X)'\Sigma_{XX}^{-1}\Sigma_{XZ}+e'_{Z|X}) \\ &= \Sigma_{XY}\Sigma_{XX}^{-1}E(X-\mu_X)(X-\mu_X)'\Sigma_{XX}^{-1}\Sigma_{XZ}+0+0+E(e_{Y|X}e'_{Z|X}) \\ &= \Sigma_{XY}\Sigma_{XX}^{-1}\Sigma_{XZ}+\Sigma_{YZ|X} \end{aligned}$$

즉, 오차항을 포함하지 않으면 마지막 항이 $\Sigma_{YZ|X}=0$이 되므로 CIA 가정을 한 것과 동일하게 추정이 된다. 그러므로 오차항을 포함하지 않는 조건부평균 대체법에 의하면 Y, Z 자료 간에 공분산이 존재한다고 하더라도 이와는 상관없이 Σ_{XY}, Σ_{XZ}의 값에 따라 일정한 값이 나오는 현상이 발생한다.

본 논문에서 고려하는 방법은 마이크로 접근법이기 때문에 대체되는 값을 조건부평균값으로 정할 뿐 오차항을 인위적으로 더해주지 않는다. 그런 이유로 유일변수와 공

통변수를 구분할 때 CIA 가정을 최대한 만족해줄 수 있도록 변수를 선택하는데 Boston Housing자료는 6개의 공통변수(X)와 수령자 파일에서 2개의 유일변수(Z), 제공자 파일에서 6개의 유일변수(Y)로 분할되었다.

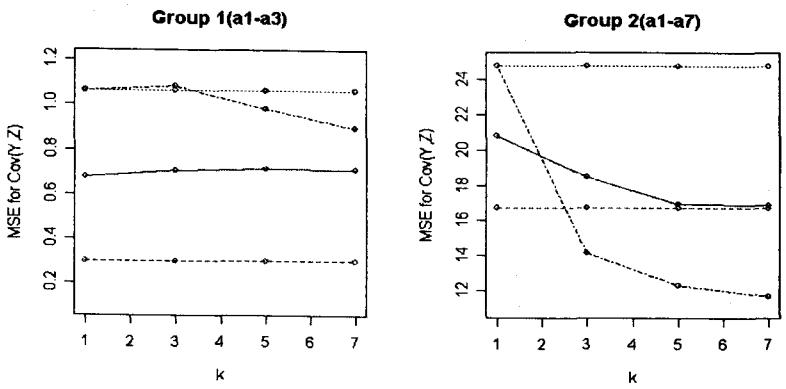
각 알고리즘의 정확성은 Y, Z의 공분산을 얼마나 정확히 추정하는가로 평가하였다. 100번의 반복적인 자료 분할에서 각 통합된 파일에서 $Cov(Y, \hat{Z})$를 추정하고 원래 값 $Cov(Y, Z)$과의 차이를 구하여 그 평균인 MSE를 다음 <표 1>에 표시하였다.

<표 1>의 "reg+k-NN" 방법은 정성석 등(2004)이 제안한 방법으로 첫 번째 행인 a3_1, a7_1은 k=1일 때 회귀식에 의한 예측치의 대체값이므로 회귀분석 알고리즘에 의하여 계산한 값과 동일하다.

이들 4가지 방법의 비교를 위하여 <그림 2>-<그림 7>에 MSE를 도시하였다. 수령자 파일의 유일변수(a1, a2, a4, a11, a14)와 a3변수와의 공분산을 추정하는데 있어서 제안한 방법이 다른 방법들에 비하여 항상 MSE가 적게 나오는 것을 확인할 수 있다. 그러나 수령자파일의 유일변수와 a7변수와의 공분산을 추정하는 경우는 항상 MSE가 작게 나오는 결과를 보이지 않았지만 다른 방법들과 비교해서 경쟁력이 있는 것으로 판단된다. 모의실험 결과에 따르면 실제 자료에서는 k-NN방법 등이 직관적으로 k가 늘어남에 따라 어느 정도 MSE가 감소하다가 증가하는 U자형을 띄어야 함에도 불구하고 계속적으로 증가하는 경우도 발생하였다. 이러한 경우 제안한 방법이 이를 보완해주는 해결책으로 제시될 수 있을 것이다.

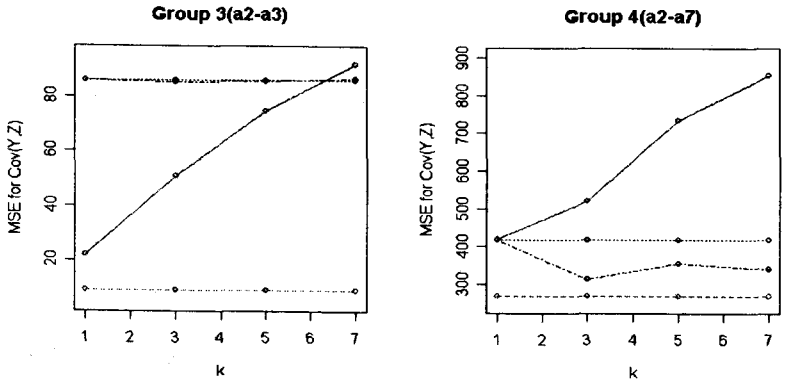
<표 1> $\{Cov(Y,Z) - Cov(Y,\hat{Z})\}$의 MSE

방법	공통변수	z=a3				z=a7			
		k=1	k=3	k=5	k=7	k=1	k=3	k=5	k=7
k-NN	a1	0.676	0.705	0.711	0.710	20.821	18.511	16.925	16.689
	a2	21.894	50.992	74.645	91.992	419.348	522.417	735.035	855.750
	a4	0.002	0.003	0.003	0.0030	0.025	0.014	0.012	0.014
	a11	0.164	0.259	0.375	0.449	2.31123	1.581	1.561	1.589
	a12	62.759	60.924	54.793	50.675	3760.16	3304.19	3000.04	2901.48
	a14	2.161	2.666	3.198	3.371	77.771	60.189	60.170	62.553
reg+k-NN (정성석 등(2004))	a1	1.063	1.084	0.982	0.894	24.776	14.125	12.292	11.641
	a2	86.043	85.356	85.660	86.600	417.626	315.286	356.136	340.247
	a4	0.004	0.004	0.003	0.003	0.022	0.013	0.011	0.010
	a11	0.244	0.239	0.251	0.256	1.886	1.107	0.979	0.861
	a12	80.237	56.314	45.486	40.881	3315.03	2634.11	2538.04	2426.70
	a14	3.330	3.280	3.491	3.698	67.659	62.999	61.018	60.532
제안 알고리즘		(h=7)		(h=20)		(h=7)		(h=20)	
	a1	0.460		0.299		25.139		16.741	
	a2	24.586		8.983		497.668		269.867	
	a4	0.002		0.002		0.035		0.030	
	a11	0.244		0.176		2.378		1.995	
	a12	50.937		32.372		3272.150		2521.980	
	a14	1.649		1.445		95.540		58.558	



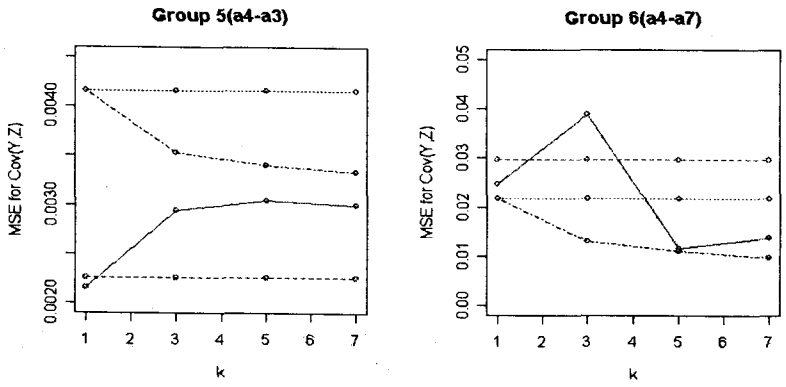
<그림 2> a1(범죄율)과 (a3(도심내 비소매업용지 비율), a7(1940년 이전에 지어진 소유가구 비율))의 공분산 추정값의 MSE

.....:제안법 ———:k-NN .....:회귀분석 —·—·—:reg+k-NN



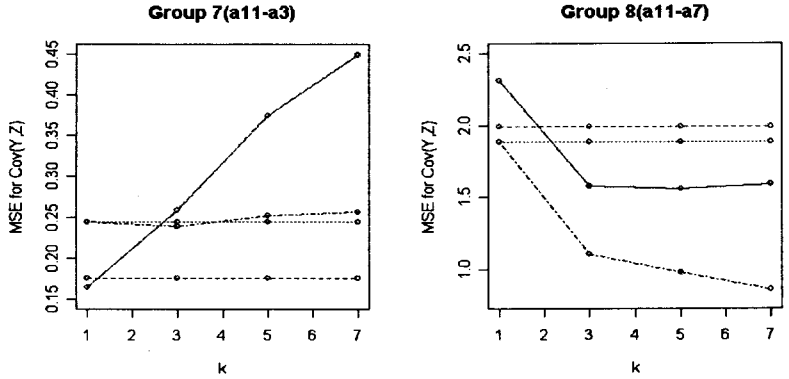
<그림 3> a2(대형 주거지역 비율)과 (a3, a7)의 공분산 추정값의 MSE

.....:제안법 ———:k-NN .....:회귀분석 —·—·—:reg+k-NN



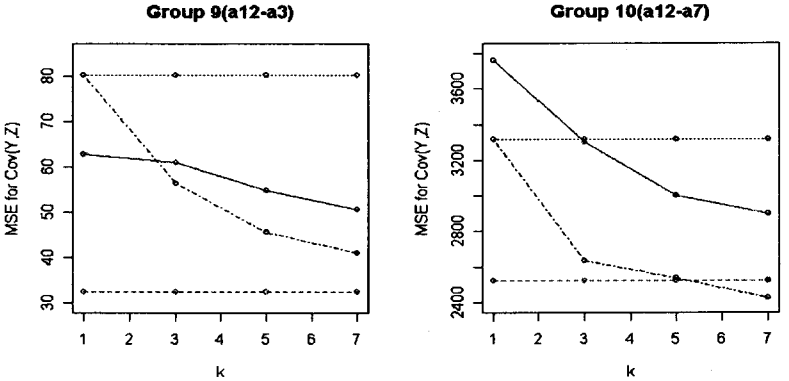
<그림 4> a4(찰스강 더미변수)와 (a3, a7)의 공분산 추정값의 MSE

.....:제안법 ———:k-NN .....:회귀분석 —·—·—:reg+k-NN



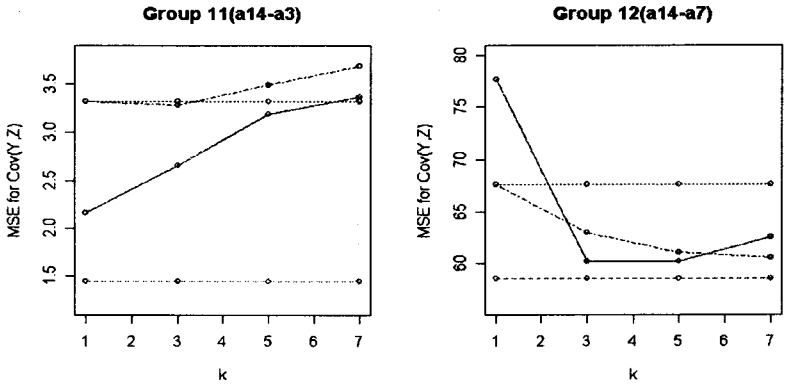
<그림 5> a11(학생 대 교사 비율)과 (a3, a7)의 공분산 추정값의 MSE

:제안법 :k-NN :회귀분석 :reg+k-NN



<그림 6> a12(흑인비율 함수)과 (a3, a7)의 공분산 추정값의 MSE

:제안법 :k-NN :회귀분석 :reg+k-NN



<그림 7> a14(소유자거주주택의 가격 중앙값)과 (a3, a7)의 공분산 추정값의 MSE

:제안법 :k-NN :회귀분석 :reg+k-NN

4. 결론

본 논문에서는 공통변수의 정보를 충분히 이용하는 h-NN후 회귀분석법에 의한 통계적 매칭방법을 제안하였다. 실제 현실 자료에서는 일반적인 k-NN 방법에서 k를 늘려감에도 그 정확성이 오히려 떨어지는 경우가 발생할 수 있고, 회귀분석방법은 낮은 적합도로 인하여 전체 평균으로 대체하는 방법과 별반 차이가 없을 수도 있다. 그래서 우리는 두 방법의 장점을 포함하고 향상된 컴퓨터 계산능력을 활용하는 h-NN후 회귀분석법을 제안하였는데, 실제 자료에 적용한 모의실험 결과 기존의 방법들과 비교하여 충분히 우수성을 견줄 수 있다는 것을 확인할 수 있었다. 단지 공통변수의 수가 많을 때 제공자파일에서 회귀식 추정에 사용할 개체를 많이 선택해야 한다는 약점이 있다. 그러나 D'Orazio et al.(2006)의 실증연구에 따르면 통계적 매칭 때 공통변수의 수가 많을수록 예측력이 증가할 거라는 예상은 들어맞지 않았다. 오히려 적은 수의 공통변수를 선택하였을 때 예측력이 증가할 수 있다는 연구결과는 우리가 제안한 방법의 단점이 해소될 수 있다는 것을 의미했다. 이는 공통변수로 사용할 수 있는 변수의 수가 많다고 공통변수의 차원을 무조건 늘리기 보다는 예측력에 도움을 주는 공통변수 만을 부분적으로 선택하는 연구가 이루어진다면 제안한 방법이 공통변수의 차원 때문에 컴퓨팅 시간이 늘어나는 문제도 해결할 수 있을 것이라 생각된다.

참고문헌

1. 정성석, 김순영, 김현진 (2004). 데이터 보강을 위한 데이터 통합기법에 관한 연구, 「응용통계연구」, 제17권, 3호, 605-617.
2. Cheng, P. E. and Chu, C. K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statistica Sinca*, 6, 63-78.
3. D'Orazio, M., Zio, M. D. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*, John Wiley & Sons.
4. Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.
5. Härdle, W. (1992). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
6. Ingram, D., O' Hare, J., Scheuren, F. and Turek, J. (2000). Statistical matching: a new validation case study, *Proceedings of the Survey Research Methods Section*, American Statistical Association.
7. Kadane, J. B. (1978). Some statistical problems in merging data files. *In Department of Treasury, Compendium of Tax Research*, pp. 159-179, Washington, DC: US Government Printing Office.
8. Moriarity, C. and Scheuren, F. (2001). Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17, 407-422.
9. Moriarity, C. and Scheuren, F. (2003). A note on Rubin's statistical

matching using file concatenation with adjusted weights and multiple imputation, *Journal of Business and Economic Statistics*, 21, 65-73.
10. Nielson, S. F. (2001) Nonparametric conditional mean imputation, *Journal of Statistical Planning and Inference*, 99, 129-150.
11. Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical *Applications and Alternative Bayesian Approaches*. New York: Springer-Verlag.
12. Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics*, 4, 87-94
13. Singh, A. C., Mantel, H., Kinack, M. and Rowe, G. (1993). Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19, 69-79.
14. van der Putten, P., Joost N. K. and Gupta, A. (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.
15. Yoshizoe, Y. and Araki, M. (1999). Use of statistical matching for household surveys in Japan, *In 52nd Session of the International Statistical Institute*, Helsinki, Finland.

[2007년 10월 접수, 2007년 10월 채택]