

산업별 에너지 소비량 자료의 결측치 추정*

권오상** · 박윤선*** · 이상준****

요 약

산업별 에너지 소비량 자료가 에너지 관련 연구 및 정책개발의 기초 자료로 사용되고 있음에도 불구하고 아직은 주요 자료가 집계된 연도가 매우 적을 뿐 아니라 결측치를 가지는 경우도 있어 자료 이용에 한계가 있다는 문제점이 있다. 본 연구에서는 결측치를 추정하는 여러 가지 방법 중에서 다중치환법(MI)을 이용하여 이러한 결측치를 메우는 분석을 시행하였다. 다중치환법에서는 에너지 소비량 자료와 더불어 산업별 생산지수를 포함하여 후자가 에너지 소비량 결측치의 예측치를 찾는 데 기여하도록 하였다. 두 가지 자료원의 일관성을 확보하기 위해 산업분류를 재구성하고, 에너지 소비량 예측치는 물론 그 분포도 도출하여 제시하였다.

주요 단어 : 산업별 에너지 소비(Industrial Energy Consumption), 결측치(Missing Data), 다중치환법(Multiple Imputation)

경제학문헌목록 주제분류 : Q4, C5, Y1

* 본 논문은 에너지경제연구원이 지원한 연구사업 「에너지 소비 데이터의 결측치 추정을 위한 모형 연구」의 성과이며, 연구보고서 「에너지 연소부문 온실가스 인벤토리 작성 및 품질개선」(2015)의 일부 내용을 발췌하여 수정·보완·확장한 논문입니다.

** 서울대학교 농경제사회학부 농업·자원경제학전공 교수 겸 농업생명과학연구원 겸무연구원(주저자), koksang@snu.ac.kr

*** 서울대학교 농경제사회학부 농업·자원경제학전공 대학원 박사과정(공동저자)

**** 에너지경제연구원 부연구위원 (공동저자)

I. 서 론

에너지 소비량 자료는 에너지 관련 연구와 정책개발의 기초자료로 사용되고 있지만, 주요 자료가 몇 가지 이유로 인해 결측치를 가지는 경우가 있다. 결측치가 발생하는 가장 큰 원인은 서베이 자료들처럼 자료자체가 매년 혹은 매기 관측되는 것이 아니고, 정기적으로 혹은 부정기적으로 특정 기간에 대해서만 조사가 진행되기 때문이다. 『에너지총조사』에서 산업부문 업종별 에너지 소비량 자료가 그 한 예가 된다. 또한 에너지 사용에 따른 대기오염 배출량 혹은 오염도 측정치처럼 실측되는 자료들 중에는 기기의 작동불량이나 다른 요인에 의해 특정 시점에 대해서는 실측치가 제시되지 않는 경우가 있으며, 특정 시설이나 측정설비의 설치가 다른 시설이나 측정설비에 비해 늦게 이루어지면서 해당 시설이나 설비로부터의 배출량 자료가 구축되지 못한 경우도 있다.

특정 연도나 기간에 중요 자료가 결측치를 가질 경우 자료 이용에 있어 여러 애로점이 나타난다. 우선 자료 값 자체가 각종 정책 판단 등에 필요한 정보를 제공하는데, 실측이 이루어지지 않은 연도는 그와 같은 정보가 공급되지 않기 때문에 발생하는 불편함이 있다. 또한 결측치를 포함하는 에너지 소비량 자료와 산업별 생산량 자료, 온실가스 배출량 자료처럼 여러 가지 통계자료 간의 관계를 계량 분석하고자 할 때 문제가 발생한다. 이 때 만약 산업별 에너지 소비량자료가 다수의 결측치를 가진다면 결측치가 발생한 연도에 대해서는 조사가 된 다른 자료들도 함께 사용할 수가 없고, 따라서 통계분석을 위한 충분한 수의 관측치를 갖추지 못해 분석의 효율성과 신뢰도가 저하될 수 있다.

이러한 문제를 해결하기 위해서는 적절한 방법을 사용해 결측치를 적합한

변수로 대체하는 절차가 필요한데, 이 때 사용할 수 있는 방법은 여러 가지가 있지만 이 중에서도 가능한 한 통계적으로 신뢰할 수 있는 방법을 사용해야 한다. 결측치가 있을 때 흔히 쉽게 사용하는 방법은 시계열적 경향성을 가지지 않는 자료라면 실측치들의 평균치를 결측치로 반영하고, 시계열적 경향성을 가진다면 결측치의 인접한 두 자료를 이용해 결측치 값을 구해내는 보간법(interpolation) 등이 있다. 그러나 이들 방법은 통계학적인 근거가 없고 자료가 가지는 통계적 불확실성을 고려하지 않으며, 결측치의 예측치를 구함에 있어 이용가능한 모든 정보를 충분히 활용한다고 보기도 어려운 방법들이다.

본 연구에서는 국가 에너지 및 온실가스 배출량 산정 등에 있어 핵심자료로 사용될 산업별 에너지 사용량 자료가 몇 개 연도에 있어 결측치를 가짐에 주목하고, 이상 설명한 바와 같이 상당한 자의성을 가지는 보간법보다는 통계적인 근거를 가지는 분석법을 검토·활용하여 이 결측치들을 구하는 작업을 수행하고자 한다. 본고가 사용하는 분석법은 통계학 분야에서 자료치환법(imputation)이라는 이름으로 1970년대 중반 이래 발전해온 기법이고, 그중에서도 다중치환법(multiple imputation, MI)이라 불리는 기법이다. 다중치환법은 통계자료의 결측치를 대체 값으로 치환하여 모두 메워주되 두 개 이상의 변수를 활용한다는 특징을 가지는데, 매 관측 시점에 있어 이들 변수 중 최소한 하나는 실측치를 가지고 있는 경우에 사용된다. 따라서 실측치가 있는 변수가 결측치를 가지는 변수의 결측치를 대체할 값을 찾는 단서를 제공하게 된다. 자료치환법에도 여러 종류가 있으며, 이들 방법은 모두 통계학적인 근거를 가지고 개발되어온 방법들이다. 자료치환법은 원래 주로 설문조사나 서베이 자료에서 모든 항목이 응답되지 않을 경우 무응답항목을 포함하는 조사표 자체를 분석에서 제거해야만 하는 손실을 막기 위해 개발되었지만, 에너지 소비 자료와 같이 주기적 혹은 비정기적으로 결측치를 가지는 자료에 대해서도 적용될 수가 있을 것이다.

우리나라의 산업별 에너지 소비량 자료는 국가 공공기관에 의해 비교적 최근 연도에 집계되기 시작하여 이용 가능한 실측치 자체가 최근 몇 개 연도에

한정되어 있다. 그나마 일부 연도에 있어서는 실측치가 제공되지 않고 있는데 본고는 매년 자료가 이용 가능한 산업별 생산지수와 결측치를 가지는 에너지 소비량간의 관계를 통계 분석하여, 후자가 결측된 연도에서도 그 값을 추정하여 제공하고자 한다. 본고가 사용하는 이러한 방식은 현재로서는 결측치의 수가 적어 통계적 유의성을 확보하는 것이 수월하지 않고 산업별 생산지수 외의 여타 관련 변수를 반영하는 데에도 한계가 있지만, 향후 추가적인 실측치가 쌓일 경우 더 신뢰도가 높은 수치를 제공할 수가 있을 것이다.

본고는 제2장에서 분석에 사용할 다중치환법(MI)의 적용 절차에 대해 설명하며, 제3장에서는 분석에 사용된 자료와 분석결과를 설명한다. 그리고 마지막 제4장에서는 분석결과를 요약하고 시사점을 찾는다.

II. 분석모형

결측치의 예측치를 구하는 방법을 선택하기 위해서는 우선 결측치가 분포되어 있는 구조와 그 결측치가 발생하는 통계적 메커니즘을 이해해야 한다. 결측치는 분포 구조에 따라 크게 단일변수형(univariate pattern), 단위무응답형(unit nonresponse pattern), 단조형(monotone pattern), 일반형(general pattern), 설계형(planned missing pattern)으로 구분된다(Little and Rubin, 2002)¹⁾. 단일변수형은 여러 변수 중 한 값만 결측치를 가지는 경우, 단위무응답형은 일부 관측, 일부 결측을 나타내는 경우, 단조형은 패널자료에서 데이터의 연속성이 끊기는 경우를 의미한다. 일반형은 앞의 세 경우에 비해 일반적이고 특정 패턴을 적용하기 어려운 분포를 나타내고, 설계형 패턴은 설문조사에서 응

1) 본절에서 설명하는 다중치환법의 기본적인 분석절차는 Little and Rubin(2002)과 Enders(2010), Horton and Kleinman(2007)을 참조하여 작성하였다. 본고가 사용할 분석 절차에 관한 보다 자세한 설명을 위해서는 김성균 외(2015)를 참조할 수 있다.

답자별로 질문 항목을 달리하는 경우 발생한다.

이상 모든 결측치 분포 구조에 있어서 중요한 것은 변수들 간에는 결측치가 발생하여도 나머지 변수 중 실측치가 최소 한 값 이상 존재한다면, 이 실측값을 이용해 결측 변수의 값을 예측하거나 파악할 수 있다는 것이다. 본 연구가 관심을 가지는 에너지 소비 자료는 산업별 제조업 생산지수와 밀접한 관련이 있을 것으로 판단된다. 이 두 변수 간의 함수 관계를 회귀 분석 등의 계량적인 방법을 통해 파악할 수 있고, 그 결과를 활용하는 다중치환법을 통해 에너지 소비량 예측치를 도출할 수 있다²⁾.

예를 들면 1차 금속 제조업의 생산지수와 에너지 소비량 자료가 <표 1>과 같이 집계되어 있다. 에너지 소비량 조사는 에너지경제연구원에 의해 1981년부터 매 3년마다 수행되고 있다. 그러나 2001년까지는 표본조사로 조사가 시행되었고, 2004년부터 한국에너지공단의 전수조사 자료를 이용하고 있다. 2004년~2009년 동안은 산업, 가정, 수송 순으로 주기적으로 에너지소비량 조사가 시행되었기 때문에 2008년, 2009년 산업 부문 자료는 누락되었고, 2005년과 2006년 역시 표본조사가 실시되었기 때문에 다른 연도에 비하면 자료가 누락된 해들이다. 본고는 이렇게 국가기관 혹은 담당 공공기관에서 공식적인 소비량 자료가 발표되지 않은 경우를 에너지 소비량이 결측된 것으로 간주한다.

이러한 <표 1>의 두 변수가 가지는 구조는 앞에서 언급된 다섯 가지의 결측치 구조 중 두 번째인 단위무응답형에 가장 가깝다고 볼 수 있다. 즉 생산지수는 통계청에서 오래전부터 정기적으로 조사해오고 있는 것으로서 매 측정시점에서 실측치가 있다. 그러나 업종별 에너지 소비량은 실측에 소요되는 비용이나 여타 사정 등으로 인해 매년 이용가능하지 않다.

2) 에너지 소비량과 상관관계를 가지는 다수의 변수를 활용할 수도 있고, 이들 변수들도 완비되거나 아니면 결측치를 가질 수 있다. 결측치의 대체 값을 찾는 이하 모든 내용은 일부 결측치를 가지는 3개 이상 다수의 변수간의 관계에 대해서도 그대로 적용된다.

<표 1> 1차 금속 제조업의 에너지 소비량과 생산지수

연 도	생산지수	에너지 소비량 (단위: 천 TOE)	실측여부
2000	70.8		0
2001	71.7		0
2002	75.3		0
2003	79.2		0
2004	83.3	18,717	1
2005	83.5		0
2006	86.6		0
2007	90.5	33,699	1
2008	91.9		0
2009	83.3		0
2010	100.0	23,823	1
2011	106.2	29,138	1
2012	106.8	28,901	1
2013	106.0	28,521	1
2014	109.9	31,664	1

우리가 얻는 자료상의 변수들을 Y 라 했을 때, 모든 값이 다 실측된 완비 자료(Y_{com})는 관측되는 자료(Y_{obs})와 결측치를 가지는 경우(Y_{mis})로 구분된다. 즉, $Y_{com} = (Y_{obs}, Y_{mis})$ 와 같이 표현할 수 있는데, <표 1>의 1차 금속 제조업을 예로 들면 2008년의 Y_{obs} 는 생산지수이고 Y_{mis} 는 에너지 소비량 자료이다.

결측치 발생 메커니즘과 관련하여, 결측치가 발생하는지의 여부가 결측치를 가지는 변수(Y_{mis}) 값 자체에 의존하는지가 특히 결측치 분석에서 중요한 의미를 가지는데, Rubin(1976)에 따르면 결측치 값 자체가 결측이 되는지의 여부에 직접 영향을 미치지 않는다면 Y_{obs} 와 Y_{mis} 의 관계만을 이용해 후자의 결측치를 추정해낼 수 있다. 반면 에너지소비량이 특정 수준을 넘어서는 경우에만 실측이 이루어지는 것처럼 결측 여부가 Y_{mis} 에 의존하는 경우에는 Y_{obs} 와 Y_{mis} 의 관계는 물론이고 결측이 발생하는 메커니즘도 함께 분석을

해야 하는 난점이 있다. 본 연구에서 사용하는 에너지 소비량 자료는 에너지 소비량에 관계없이 특정 연도에만 결측이 발생한 것이므로 Rubin(1976)의 조건을 충족한다고 할 수 있다.

통상적으로 결측치를 메울 때에는 통계적 근거 없이 일종의 보간법을 사용한다. 자료가 시계열적 경향성을 가지지 않는 경우 실측치의 평균을 해당 변수의 모든 결측치에 대입할 수 있다. 그러나 이 방법은 본 연구에서 사용하는 에너지 소비량 자료가 시간이 지나면서 대체로 증가하거나 감소하는 경향성을 가지기 때문에 적합한 방법이 될 수 없다. 또 다른 방법으로서 스플라인(spline)보간법이 있는데, 이는 결측치의 앞 뒤 두 관측치를 선형으로 연결하고 그 가운데 값을 결측치의 대체 값으로 간주하는 방법이다. 또는 세 점 이상의 관측치를 2차 함수형태로 연결하고 이를 이용해 결측치 값을 구할 수도 있다. 이와 같은 보간법은 자료가 가질 수 있는 확률적 영향은 고려하지 않으며, <표 1>의 생산지수처럼 전체 관측치에서 실측되는 관련변수를 모두 활용하지도 않는다는 점에서 비효율적인 방법이다.

방법 자체는 통계적이지만, 보간법의 한 형태로 간주할 수 있는 것이 모든 변수가 실측이 된 관측치들만을 가지고 보조변수(auxiliary variables)를 이용해 결측치의 예측치를 구하는 방법이다. 예를 들면 다음과 같은 선형회귀식을 통해 결측치를 추정할 수 있다.

$$(1) \quad Y_{mis,t} = \alpha + \beta Y_{obs,t} + \epsilon_t$$

식 (1)에서 ϵ_t 는 교란항인데, 위의 회귀식은 두 변수 모두 관측되는 연도의 자료만을 이용해 추정된다. 이러한 회귀분석을 이용한 보간법은 유용한 방법이지만 자료가 결측된 경우 두 변수 간에 100%의 상관관계가 있다고 가정하는 방법이다. 이는 실제로 성립하기 어렵기 때문에 전체 표본에 있어서 두 변수간의 상관관계를 과장할 수 있다.

기존 방법의 단점을 보완하고자 본고가 사용하는 자료치환법은 크게 단일

치환법(single imputation)과 다중치환법(multiple imputation)으로 구분되는데, 첫 번째 방법은 앞의 회귀분석을 이용한 예측치처럼 각 관측치의 결측치를 대체할 하나의 숫자만을 구하는 방법이고, 두 번째 방법은 단일치환법이 사용하는 분석법을 확대 적용하여 각 결측치를 대체할 다수의 대체 값을 구해내는 방법이다. 결측치에 관한 현대적 통계분석기법은 모두 다중치환법(MI)을 표준적인 분석법으로 하고 있고, 본고 역시 이 방법을 사용한다.

다중치환법의 목적은 단순히 결측치를 치환하는 데 있는 것이 아니고 치환을 거쳐 완비된 자료를 이용해 그 다음 단계의 통계분석을 행하는 데 있다³⁾. 따라서 이 방법은 크게 다음의 세 가지 단계를 가진다.

- i) 치환단계: 결측치를 포함할 수도 있는 다수의 변수로 구성된 자료에 치환기법을 적용해 각 결측치별로 다수의 대체 값을 생성해낸다. 각 결측치의 대체 값을 M 회 구했다면, M 가지의 완비된 표본이 생성된다.
- ii) 개별분석단계: 치환된 자료를 포함하여 완비된 M 개의 표본을 각각 독립적으로 이용해 M 회의 원하는 분석(예: 회귀분석)을 한다.
- iii) 통합단계: 단계 ii)에서 이루어진 M 회의 분석결과를 모두 통합하여 변수들의 특성이나 변수들 간 관계에 대한 원하는 정보를 도출한다.

다중치환법을 적용하기 위한 사전 단계로 먼저 단일치환법을 적용한다. 이를 위해 <표 1>의 두 변수처럼 우리가 고려하는 변수를 n 가지로 일반화하고, n -벡터를 Y , 그 평균의 벡터를 μ , 분산공분산행렬을 Σ 라 하자. 이 때 각 관측치에서의 로그우도 값을 모두 더한 값을 극대화하는 Σ 와 $\mu(=\hat{\Sigma}, \hat{\mu})$ 를 추정하고 결측치의 예측치까지 도출하는 최우추정법(MLE)이 사용될 수 있다. 그러나 실제 분석에 있어서는 로그우도함수를 바로 극대화하는 최우추정법보다는 EM 알고리즘을 이용한 반복 추정법이 더 효율적이기 때문에 대

3) Honaker and King(2010)의 연구는 자료치환법을 여기 저기 흠이 페인 스위스 치즈의 흠을 모두 메워 매끈한 치즈로 만드는 과정에 비유하고 있다.

부분의 연구가 이를 채용하고 있다. 위의 예에 대해 EM 알고리즘을 적용하면, 먼저 최초의 Σ 와 μ 의 값을 가정한 상태에서 각 관측치의 실측 자료들($= Y_{obs}$)을 이용해 결측치($= Y_{mis}$)의 예측치를 구해낸다(E-step). 이렇게 결측치가 메워져 완비된 자료를 MLE기법을 통해 파라미터 Σ 와 μ 의 새로운 추정치를 도출하는데 사용한다(M-step). 파라미터 추정치가 더 이상 변하지 않을 때까지 이 절차를 반복하여 최종 추정치 $\hat{\Sigma}$ 와 $\hat{\mu}$ 를 구한다.

이어서 진행되는 다중치환법은 각 결측치에 대해 다수의 대체 값을 구해낸다. 이렇게 하기 위해서는 다수의 파라미터 추정치가 필요하다. 다시 말해, 만약 결측치별로 M 개의 대체 값을 구해내려면 추정치 ($\hat{\Sigma}$, $\hat{\mu}$)의 조합이 M 개가 필요하다. 앞에서 보았던 최우추정법이나 EM 알고리즘은 우리가 가진 자료에 대해서는 하나의 추정치 조합만을 제공하기 때문에, 기존의 다중치환법 문헌들에서는 베이즈(Bayes) 추정법을 적용하여 추정 파라미터 Σ 와 μ 의 값이 아닌 그 분포를 추정한다⁴⁾.

베이즈 추정법을 통해 M 조합의 완비된 자료가 얻어진 후, 각 조합에서의 파라미터의 평균치나 대체 값의 평균치를 구할 수 있다.

$$(2) \quad \bar{\theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m, \quad \bar{Y}_{mis} = \frac{1}{M} \sum_{m=1}^M \hat{Y}_{mis,m}$$

아울러 각 조합에서의 파라미터 $\hat{\theta}_m$ 과 그 전체 평균 간의 차이를 이용해 도출한 다음의 $\bar{V}$ 를 파라미터의 분산으로 간주한다⁵⁾.

4) 대안적 방법으로서 최우추정법 등을 통해 얻어진 추정치 ($\hat{\Sigma}$, $\hat{\mu}$)의 분산공분산행렬을 도출한 후, 이들 추정치가 추정된 분산공분산행렬을 따르는 다항정규분포를 가진다고 보고 이들 추정치의 조합을 몬테카를로(Monte Carlo) 시뮬레이션을 이용해 추출하고, 추출된 각 조합에서 결측치의 대체 값을 모두 계산하여 대체 값을 시뮬레이션 횟수만큼 확보하는 방법을 사용할 수도 있다.

5) 그 외 M 개의 완비된 자료 각각에서의 각 파라미터의 표준편차를 구해 그 제곱 합을

$$(3) \quad \bar{V} = \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_m - \bar{\theta})^2$$

Ⅲ. 분석결과

본 연구에서는 다중치환법을 이용하여 제조업분야 에너지원별 소비량 데이터의 결측 값을 추정하고자 하였다. 앞서 언급하였듯이 제조업분야 에너지 소비량은 에너지경제연구원에 의해 1981년부터 매 3년마다 수행되고 있다. 그러나 2001년까지는 표본조사로 조사가 시행되었고, 2004년부터 매년 시행되는 한국에너지공단의 전수조사 자료를 이용하고 있다. 한국에너지공단에서 운영하는 국가온실가스 배출량 종합정보 시스템(NETIS)의 「산업부문(제조업) 에너지원별 소비량」은 2004년~2009년 동안 산업, 가정, 수송 순으로 주기적으로 에너지소비량 조사가 시행되었기 때문에 2008년, 2009년 산업 부문 자료는 누락되었다. 또한 2005년, 2006년 자료는 전수조사가 아닌 표본조사 자료이므로 타 연도의 전수조사 자료와 성격이 다를 수 있기 때문에 결측치로 두고 분석을 진행하였다. 따라서 본 연구에서는 NETIS의 2004년, 2007년, 2010년, 2011년, 2012년, 2013년, 2014년 산업부문 에너지원별 소비량 자료를 사용하였다.

2004년, 2007년 데이터는 제조업 분야를 총 23개로 구분하여 에너지원별 소비량을 제시하고 있다. 그러나 2008년 제9차 한국표준산업분류 개정 분류가 시행되면서 2010년 이후의 자료와 제조업 분류에서 차이를 보이고 있다⁶⁾. 또

구한 후, M 으로 나누어 준 소위 치환 내 분산(within-imputation variance)을 구해 이를 추가적인 분산자료로 더해주기도 한다(Enders 2010, pp. 222-223). 그러나 치환 내 분산은 실측치까지도 포함되어 계산되며, 본고는 다중치환을 시행함에 있어 결측치 추정치의 변이에 주로 관심을 가지고 있기 때문에 식 (3)의 치환 간 분산(between-imputation variance)을 결측치 대체 값의 변이로 간주한다.

6) 제9차 한국표준산업분류 개정에 따르면 재생원료 생산업이 C.제조업(10~33)에서 E.하

한 2012년부터는 한국표준산업분류의 소분류(3자리 숫자) 기준에 따라 제조업을 총 84개로 분류하여 에너지 소비량 데이터를 제공하고 있다. 2012년 이후의 산업분류를 기준으로 에너지 결측치를 추정할 경우 2개의 관측치를 바탕으로 10개년도 이상의 결측치를 추정해야하므로 추정의 정확도가 다소 낮을 수 있다. 따라서 본 연구에서는 한국표준산업분류 중분류(2자리 숫자) 기준에 따라 제조업을 분류한 뒤 분석을 진행하고자 한다. 중분류에 따르면 제조업은 23개로 구분되는데 이 중 21.의료용 물질 및 의약품 제조업은 제8차 한국표준산업분류에 따라 화학물질 및 화학제품 제조업에 포함시켰고, 32.가구 제조업의 경우 가구 및 기타 제품 제조업에 포함시켰다. 또한 2004년, 2007년 데이터를 한국표준산업분류 기준에 맞추기 위해 전자장비 제조업, 컴퓨터 및 사무용기기 제조업 분야를 26.전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업 분야로 설정하였다. 본고가 취하는 이러한 산업구분결과는 <표 2>에 정리되어 있다.

제조업 분야 에너지 소비 데이터 결측치 추정에 사용되는 Y_{obs} 변수로는 본고에서는 통계청에서 제공하는 2000년~2015년 산업별 광공업생산지수를 이용하였다. 식·음료품 제조업은 식료품, 음료제조업으로, 가구 및 기타 산업은 가구, 기타제품 제조업으로 구분되므로 각각의 생산지수는 두 산업의 해당년도 생산액 비중을 가중 평균하였다.

실측이 된 21개 제조업 산업별 에너지 사용량 자료는 <표 3>에 정리되어 있다. 본고는 각 제조업의 생산지수를 이용해 표본 조사가 이루어진 2000년, 2001년, 2002년, 2003년, 2005년, 2006년, 자료가 결측된 2008년, 2009년, 그리고 에너지 사용량 자료가 집계되기 이전인 2015년의 총 9개 연도의 결측치를 구해내고자 한다.

수·폐기물 처리, 원료재생 및 환경복원업(37~79)으로 이동하였고, 출판업이 J.출판, 영상, 방송통신 및 정보서비스업(58~63) 분류로 이동하였다.

<표 2> 한국표준산업분류에 따른 제조업 산업분류

에너지 소비량 자료 상 제조업 산업분류						최종 산업분류
2004	2007	2010	2012	2013	2014	
음·식료품 제조업		음·식료품 제조업				1. 음·식료품 제조업
담배 제조업		담배 제조업				2. 담배 제조업
섬유제품 제조업		섬유제품 제조업				3. 섬유제품 제조업
의류제품 제조업		의복, 의복액세서리 및 모피제품 제조업				4. 의복, 의복액세서리 및 모피제품 제조업
가죽제품 제조업		가죽, 가방 및 신발 제조업				5. 가죽, 가방 및 신발
나무제품 제조업		목재 및 나무제품 제조업				6. 목재 및 나무제품
종이제품 제조업		펄프, 종이 및 종이제품 제조업				7. 펄프, 종이 및 종이제품
출판업		인쇄 및 기록매체 복제업				8. 인쇄 및 기록매체 복제업
석유제품 제조업		코크스, 연탄 및 석유정제품 제조업				9. 코크스, 연탄 및 석유정제품 제조업
화학제품 제조업		화학물질/화학제품 의료용 물질 및 의약품 제조업				10. 화학물질/화학제품
고무 및 플라스틱 제품 제조업		고무제품 및 플라스틱제품 제조업				11. 고무 및 플라스틱
비금속 광물 제품 제조업		비금속 광물제품 제조업				12. 비금속 광물제품
제 1차 금속산업		1차 금속 제조업				13. 1차 금속 제조업
조립금속제품 제조업		금속가공제품 제조업				14. 금속가공제품
기타 전기기계 및 장비 제조업		기타 기계 및 장비 제조업				15. 기타 기계 및 장비
컴퓨터 및 사무용 기기 제조업		전자부품, 컴퓨터, 영상, 음향 및 통신장비				16. 전자부품, 컴퓨터, 영상, 음향 및 통신장비
전기기기, 변환장치 제조업						
전자장비 제조업		전기 장비 제조업				17. 전기 장비 제조업
의료, 정밀 광학기기 제조업		의료, 정밀, 광학기기 및 시계 제조업				18. 의료, 정밀, 광학기기 및 시계 제조업
자동차 제조업		자동차 및 트레일러 제조업				19. 자동차 및 트레일러
운송장비 제조업		기타 운송장비 제조업				20. 기타 운송장비
기타 제품 제조업		기타 제품 제조업				21. 가구 및 기타 제품 제조업
		가구 제조업				
재생재료 가공 처리업						채외

<표 3> 제조업 업종별 에너지 소비량

단위: 천TOE

연도	2004	2007	2010	2011	2012	2013	2014
합계	81,323	108,446	98,960	112,742	116,198	117,383	122,169
음·식료품	1,772	2,207	1,635	2,067	2,083	2,192	2,256
담배	30	29	28	29	33	34	25
섬유제품	1,620	1,819	1,074	1,529	1,546	1,507	1,511
의복, 의복 액세서리 및 모피제품	90	104	31	91	129	111	141
가죽, 가방 및 신발	72	82	27	60	59	81	85
목재 및 나무제품	244	461	291	559	574	483	550
펄프, 종이 및 종이 제품	2,357	2,089	2,085	2,266	2,114	2,115	2,122
인쇄 및 기록매체 복제업	113	188	63	133	143	155	157
코크스, 연탄 및 석유정제품	16,401	29,799	29,060	32,397	34,245	32,011	32,891
화학물질 및 화학 제품	25,745	22,560	26,356	28,458	28,630	31,231	31,575
고무제품 및 플라 스틱제품	1,173	1,435	1,075	276	1,697	1,887	1,985
비금속 광물제품	6,657	6,126	5,415	6,159	6,099	6,370	6,411
1차 금속	18,717	33,699	23,823	29,138	28,901	28,521	31,664
금속가공제품 제조업	854	1,313	1,376	1,298	1,541	1,780	1,861
컴퓨터, 영상, 음향 및 통신장비	1,933	2,849	3,137	3,718	3,968	4,227	4,202
의료, 정밀, 광학기기 및 시계	71	121	88	140	142	165	167
전기 장비	407	443	377	528	599	630	623
기타 기계 및 장비	1,090	985	623	1,066	988	1,031	1,152
자동차 및 트레일러	1,369	1,560	1,688	2,055	1,913	1,942	1,973
기타 운송장비	314	378	414	563	591	725	581
가구 및 기타 제품	293	200	294	214	205	186	239

자료: 국가온실가스 배출량 종합정보 시스템(NETIS) 제조업 에너지 소비 데이터를 한국
표준산업분류 중분류 기준에 따라 구분한 뒤 연도별 제조업 분류의 일치를 위해
<표 2>와 같이 재분류하였음

21개 산업 중 하나인 1차 금속제조업을 예를 들어 분석을 해보면, 그 결과는 <표 4>와 같이 정리된다. <표 4>의 1열과 2열은 1차 금속제조업의 생산지수와 에너지 사용량 실측치를 보여준다. 3열은 실측이 된 연도의 자료만을 이용해 두 변수간의 회귀분석을 한 후, 그 결과에 각 연도의 생산지수를 대입하여 도출한 에너지 사용량 예측치이다. 마지막 4열은 앞에서 설명된 다중치환법(MI)을 <표 1>의 1차 금속제조업 자료에 적용하여 도출한 각 연도에서의 대체 값의 평균치를 나타낸다⁷⁾. MI분석의 E-step을 위해 앞서 언급한 식(1)의 선형 회귀식을 가정하되, 생산지수와 에너지 소비량이 모두 0보다 작을 수 없음을 반영하여 교란항 ϵ_t 가 절단정규분포(truncated normal distribution)를 따른다고 가정하였다. M 의 값으로는 100을 사용하였고, <표 4>의 다중치환법 예측치는 도출된 100개의 대체 값의 평균이다.

분석결과를 보면, 우선 선형회귀식의 경우와 달리 다중치환법에서는 실측치가 있는 연도의 경우 자동적으로 그 실측치를 채택함을 볼 수 있다. 선형회귀 분석 보간법이 보여주는 예측치와 다중치환법이 적용된 예측치가 서로 유사한 경향을 보여주고는 있지만, 후자의 경우 앞에서 제기된 여러 가지 통계적 기준을 감안하여 도출된 수치라는 점에 있어 차이가 있다.

MI의 $M(=100)$ 개의 자료 집합을 각각 이용해 얻어진 결측치 예측치를 <표 4>의 4열처럼 구하여 결측치 대신 사용하면 총 16개의 관측치가 완비된다. 이 완비된 자료를 이용해 우리는 원하는 경제 분석을 행할 수가 있다. 예를 들면 1차 금속 제조업의 생산지수 한 단위 변화가 유발할 에너지 사용량 변화를 알고자 한다면, 총 16개의 관측치를 모두 이용하여 다음과 같이 절단분포 회귀분석을 할 수가 있다.

$$Y_t = 2287.0 + 253.9X_t, \text{ Nobs} = 16$$

$$(0.13) \quad (1.48)$$

7) 본고의 모든 분석을 위해 Stata 14을 사용하였다. MI분석을 위해서는 Stata의 mi기능을 이용하였다.

단, Y_t = t 연도의 에너지 소비량, X_t = t 연도의 생산지수

반면, 수식 (1)처럼 통상적인 회귀분석을 통해 두 변수간의 관계를 분석하고자 하되, 실측치가 있는 연도는 실측치를, 결측치가 있는 경우 통상적인 회귀분석 보간법을 이용해 보충된 자료를 Y_t 에 대입해 완비된 자료에 회귀분석을 적용하면 추정결과는 다음과 같다.

$$Y_t = 1423.9 + 262.6X_t, \quad R^2 = 0.63, \quad Nobs = 16$$

(0.29) (4.93)

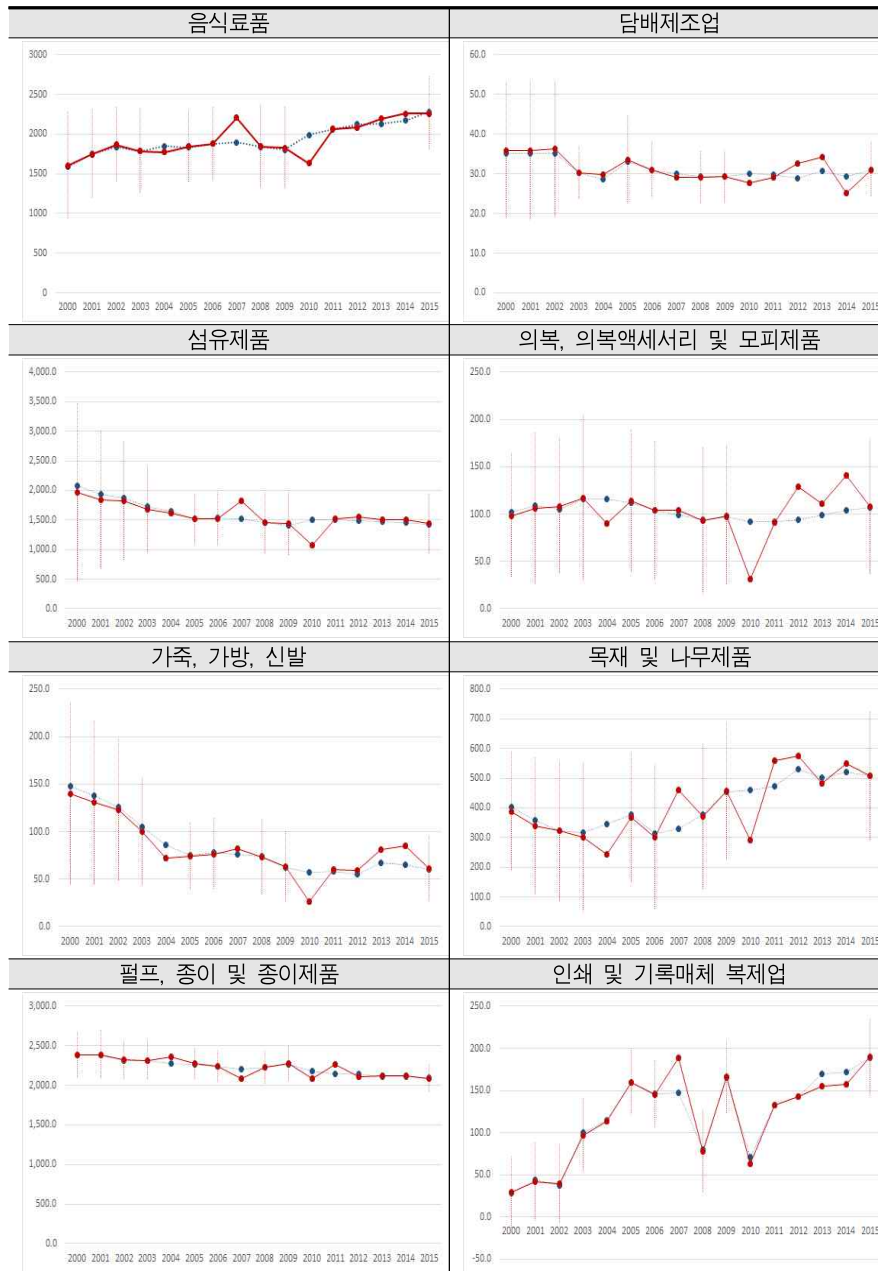
서로 다른 두 가지 방법을 이용해 얻어진 결측치를 이용한 사후 통계분석 결과는 서로 유사하다. 추정치도 그러하며, 그 통계적 유의성 역시 유사하다. 그러나 그럼에도 불구하고 통상적인 회귀분석을 이용한 선형보간법은 Y_t 가 결측된 연도에 있어서는 두 변수 Y_t 와 X_t 가 모두 추정된 하나의 선형식 위에 위치하여 두 변수간의 상관관계가 100%라는 가정을 하고 있기 때문에 이들 자료의 분산이 0으로 왜곡되고, 따라서 이렇게 구축된 자료를 이용한 통계분석결과의 R^2 값이나 여타 추정결과들도 왜곡된다는 것이 잘 알려져 있다 (Enders 2010, pp. 44-46; Beale and Little 1975).

<표 4> 다중치환법 적용결과: 1차 금속 제조업

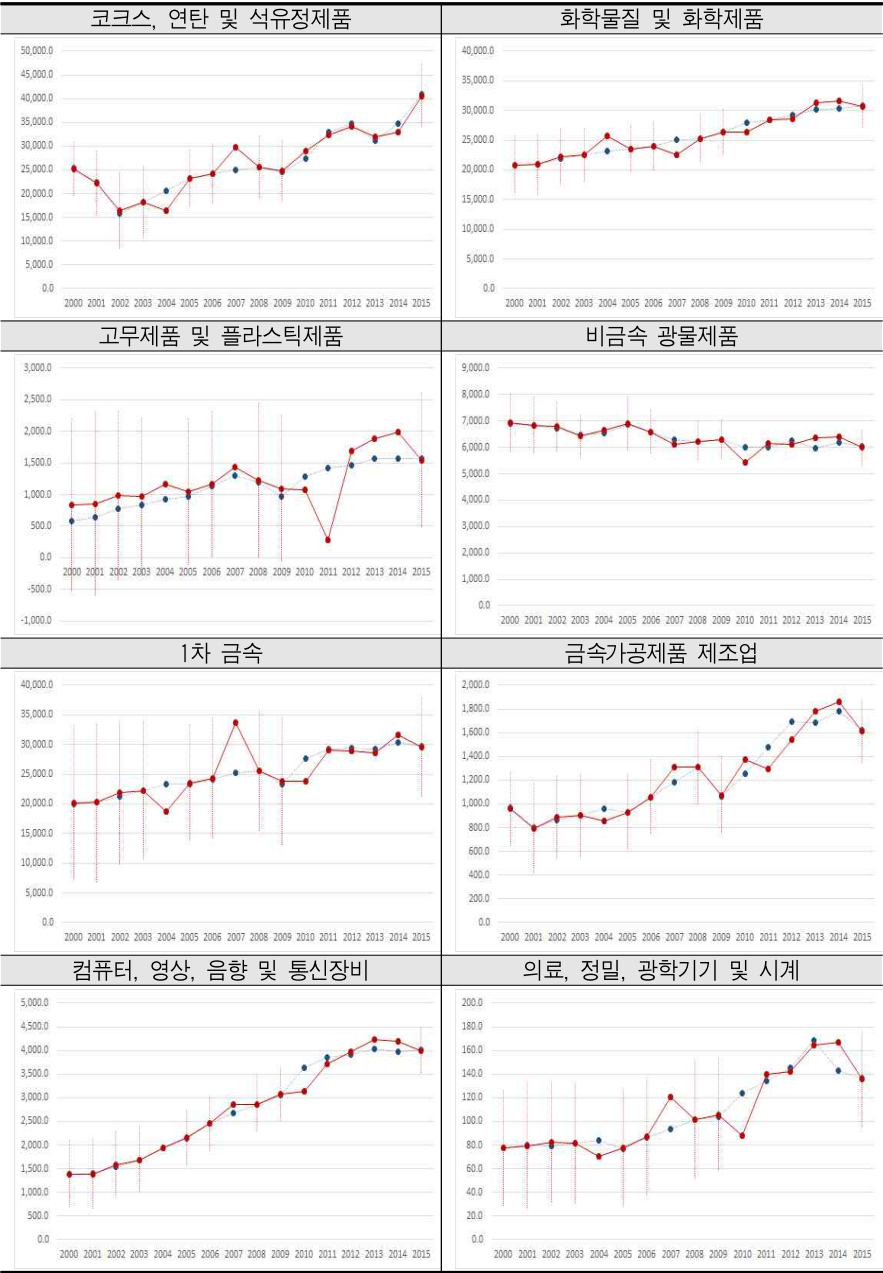
연도	생산지수	에너지 소비량	예측치: 회귀분석보간법	예측치: 다중치환법
2000	70.8		20,013	20,160
2001	71.7		20,249	20,257
2002	75.3		21,194	21,928
2003	79.2		22,218	22,275
2004	83.3	18,717	23,295	18,717
2005	83.5		23,347	23,574
2006	86.6		24,161	24,248
2007	90.5	33,699	25,185	33,699
2008	91.9		25,552	25,538
2009	83.3		23,295	23,769
2010	100.0	23,823	27,679	23,823
2011	106.2	29,138	29,307	29,138
2012	106.8	28,901	29,465	28,901
2013	106.0	28,521	29,255	28,521
2014	109.9	31,664	30,278	31,664
2015	108.0		29,780	29,512

이상의 MI분석은 모든 21개 산업에 대해 시행되었으며, 그 결과 <그림 1> 이 결측치를 포함하는 자료의 변화 형태를 보여준다. 아울러 <부표>는 모든 산업별 에너지 소비량의 실측치는 물론 MI를 통해 메워진 결측치 대체 값까지 모두 숫자로 보여준다. <그림 1>의 막대그래프는 분산 추정치를 이용해 도출한 95% 신뢰구간을 나타낸다.

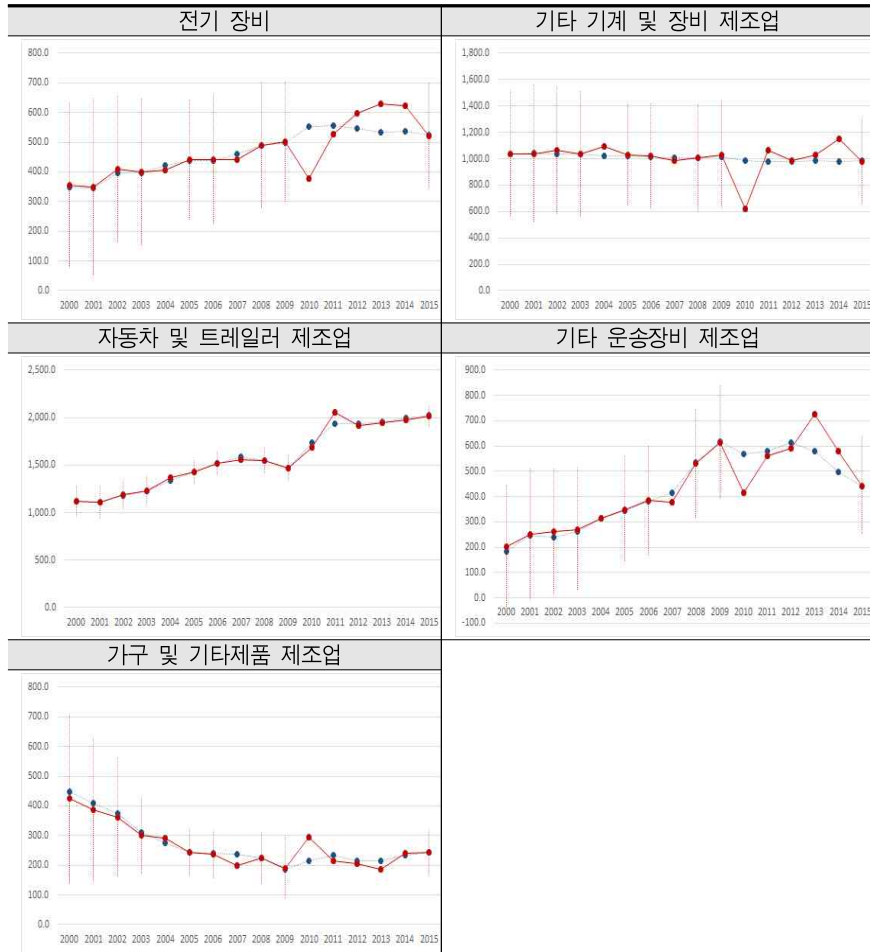
<그림 1> 제조업 분야별 다중치환법에 의한 에너지 사용량 예측치



<그림 1> 제조업 분야별 다중치환법에 의한 에너지 사용량 예측치(계속)



<그림 1> 제조업 분야별 다중치환법에 의한 에너지 사용량 예측치(계속)



주: —●— 다중치환법, 회귀보간법

에너지 소비량의 변동형태는 경기변동에 따른 등락은 있지만 컴퓨터, 영상, 음향 및 통신장비나 자동차 및 트레일러 제조업처럼 시간이 지나면서 지속적으로 증가하는 산업이 있는 반면, 섬유제품, 가죽, 가방, 신발, 기타제품 제조업처럼 지속적으로 감소하는 산업도 있다.

아울러 선형보간법 등을 사용한 경우와 달리 <그림 1>의 다중치환법을 통한 대체 값들은 인접한 실측치와 반드시 선형이거나 단조적인 관계를 가지지 않는다는 점도 확인이 된다. 특히 경기변동에 따라 산업별 생산 활동 수준이 변하면서 단기적인 에너지 소비량 변동도 관측된다.

그리고 결측치 대체 값의 95% 신뢰구간의 경우 추정치를 얻기 위한 보조 변수로 사용된 생산지수와 에너지 소비량 간의 상관관계가 매우 높은 코크스, 연탄 및 석유정제품, 화학물질 및 화학제품, 비금속 광물제품, 펄프, 종이 및 종이제품, 자동차 및 트레일러 제조업, 컴퓨터, 영상, 음향 및 통신장비의 경우 각 연도별 분산 값이 작고, t-값이 높아 MI기법을 이용한 결측치 추정치의 통계적 신뢰도가 이들 산업에서 특히 높다는 것도 확인이 된다.

반면 음식료품, 담배제조업, 의복, 의복 액세서리 및 모피제품, 목재 및 나무제품, 고무제품 및 플라스틱제품 등의 경우 결측치 예측치의 분산이 크고 생산지수의 t-통계량 값이 높지 않은데, 이들 산업의 경우 에너지 집약도가 비교적 낮은 산업 군으로 산업별 생산지수가 에너지 소비량 변화를 설명하는 변수로서 설명력이 상대적으로 낮다. 이들 산업의 경우 따라서 산업별 생산지수 외에 항상 실측이 되는 다른 설명변수를 찾아 다중치환법을 적용할 수 있지만, 적절한 변수를 실제로 찾기는 쉽지 않을 것으로 보인다. 대안으로 이들 산업별 에너지 사용량 스스로의 시계열자료만을 이용하되, 특히 칼만필터(Kalman filter) 분석법을 사용할 경우 새로운 관측치가 이용 가능할 때마다 상태변수(state variable)를 업데이트하는 기법의 특성을 활용하여 결측치의 대체 값을 찾을 수도 있다(Tsay, 2010). 현재로서는 7개 연도의 에너지 소비량 실측치만 이용 가능하여 정교한 시계열 분석법을 적용하기는 어려울 것인데, 이 경우 음식료품 등의 산업의 경우 관례적인 단순 보간법 등을 MI기법과 더불어 활용할 수도 있을 것이다.

IV. 요약 및 결론

본고는 에너지 소비량 자료가 에너지 관련 연구 및 정책개발의 기초 자료로 사용되고 있음에도 불구하고 아직은 주요 자료가 집계된 연도가 매우 적을 뿐 아니라 결측치를 가지는 경우도 있어 자료 이용에 한계가 있다는 문제점에 착안하여 이러한 결측치를 메우는 분석을 시행하였다. 특히 에너지 관련 자료 중에서 활용도가 높은 산업분야 에너지 소비량 데이터는 2004년부터 조사가 시행되었지만 일부 연도에서 조사가 시행되지 않아 결측치가 발생하였고, 이에 현재 이용 가능한 실측치가 7개 연도에 불과하다. 본 연구에서는 결측치를 추정하는 여러 가지 방법 중에서 통계적으로 근거를 가지는 자료치환법, 이 중에서도 다중치환법(MI)를 이용하여 산업별 에너지 소비 데이터를 추정하였다. 다중치환법에는 에너지 소비량 자료와 더불어 산업별 생산지수를 포함하여 후자가 에너지 소비량 결측치의 예측치를 찾는 데 기여하도록 하였다.

본고가 사용하는 다중치환법을 적용할 경우 실측치가 존재하는 연도에서는 단순회귀분석 보간법과 달리 실측치 자체를, 그리고 결측치가 있는 연도에서는 다수의 치환과정을 통해 얻어진 예측치의 평균값을 결측치의 대체 값으로 선택하기 때문에 단순회귀분석 보간법에 비해 실제 자료에 더 근접한 예측치를 추정할 수 있다. 아울러 자료가 결측된 연도에 있어서도 통상적인 회귀분석 보간법과는 달리 통계적 변이를 가지는 예측치를 제공하여, 완비된 자료를 이용한 후속 통계분석의 신뢰도를 높인다는 장점도 가진다.

본고에서 제시하는 결측치의 예측치 자체는 현재 이용 가능한 실측치의 수가 너무 적은 관계로 높은 신뢰도를 부여하기는 어렵지만, 다중치환법이 본고가 설정한 분석목적을 위해 비교적 성공적으로 활용될 수 있음이 나타난다.

국가 에너지관련 정책분석과 경제 분석에 필요한 기초자료의 결측치를 보완해야 할 필요성이 자주 대두되는 상황에서 이들 분석기법은 다양하게 활용될 수 있을 것이며, 그 적절한 적용 대상을 찾고 분석의 신뢰도 향상을 위해 필요한 관련 자료를 추가로 확보하는 노력 등도 필요할 것이다.

접수일(2016년 5월 25일), 수정일(2016년 7월 15일), 게재확정일(2016년 8월 1일)

◎ 참 고 문 헌 ◎

- 김성균·이상준·오경수·유학식·김지현·임효상·오상근·권오상·안병일·이서혜 2015, 「에너지 연소부문 온실가스 인벤토리 작성 및 품질개선」, 에너지경제연구원.
- 통계청, 광업제조업동향조사, 각 연도. <http://kosis.kr/>
- 한국에너지공단 국가온실가스 배출량 종합정보시스템(NETIS), <http://netis.kemco.or.kr/>
- Beale, E. M. L. and R. J. A. Little, 1975, Missing Values in Multivariate Analysis, *Journal of Royal Statistical Society, Series B*, 37, 129-145.
- Enders, C. K. 2010. *Applied Missing Data Analysis*, NY: The Guilford Press.
- Honaker, J. and G. King. 2010. “What to Do about Missing Values in Time-Series Cross-Section Data,” *American Journal of Political Science* 54: 561-581.
- Horton, N. J., and K. P. Kleinman. 2007. “Much Ado about Noting: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models,” *The American Statistician* 61: 79-90.
- Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.
- Rubin, D. B. 1976. “Inference and Missing Data,” *Biometrika* 63: 581-592.
- Tsay, R. S. 2010. *Analysis of Financial Time Series*. 3rd ed., NJ: Wiley.

<부표> 다중치환법을 이용한 제조업 분야별 에너지 사용량 추정값

단위: 천 TOE

연도	음식료품	담배제조업	섬유 제품	의복 및 악세서리	가죽, 가방	목재, 나무	펄프, 종이
2000	1,602	35.8	1,962	98.5	140.0	388.5	2,383
2001	1,744	35.8	1,836	106.3	130.4	339.2	2,388
2002	1,864	36.3	1,817	108.5	122.5	321.9	2,329
2003	1,784	30.2	1,677	117.1	100.4	301.1	2,315
2004	1,772	29.8	1,620	90.1	72.2	243.9	2,357
2005	1,843	33.5	1,520	113.7	74.0	369.0	2,272
2006	1,879	31.0	1,517	104.0	76.2	302.4	2,240
2007	2,207	29.0	1,819	104.2	82.2	461.1	2,089
2008	1,846	29.1	1,454	92.7	72.7	369.9	2,226
2009	1,827	29.4	1,434	98.6	62.8	456.6	2,273
2010	1,635	27.6	1,074	31.2	26.7	291.1	2,085
2011	2,067	29.2	1,529	90.8	60.2	558.5	2,266
2012	2,083	32.6	1,546	128.5	59.0	574.1	2,114
2013	2,192	34.2	1,507	110.6	81.2	483.0	2,115
2014	2,256	25.2	1,511	140.9	84.7	550.2	2,122
2015	2,257	31.0	1,437	107.9	60.9	507.7	2,091
연도	인쇄 및 기록매체	코크스, 석유정제	화학물질 및 제품	고무/ 플라스틱	비금속	금속가공	전자부품, 영상 및 통신
2000	29.6	25,101	20,825	839	6,935	958	1,387
2001	41.4	22,155	20,988	853	6,826	795	1,392
2002	39.7	16,376	22,217	987	6,781	886	1,593
2003	97.0	18,301	22,492	967	6,445	901	1,691
2004	113.1	16,401	25,745	1,173	6,657	854	1,933
2005	159.6	23,192	23,567	1,052	6,898	931	2,160
2006	145.5	24,101	23,948	1,169	6,585	1,059	2,462
2007	188.2	29,799	22,560	1,435	6,126	1,313	2,849
2008	78.1	25,547	25,310	1,223	6,224	1,310	2,862
2009	166.0	24,868	26,378	1,098	6,307	1,074	3,073
2010	62.5	29,060	26,356	1,075	5,415	1,376	3,137
2011	133.1	32,397	28,458	276	6,159	1,298	3,718
2012	142.6	34,245	28,630	1,697	6,099	1,541	3,968
2013	154.6	32,011	31,231	1,887	6,370	1,780	4,227
2014	156.9	32,891	31,575	1,985	6,411	1,861	4,202
2015	189.6	40,622	30,649	1,545	6,012	1,612	4,002

산업별 에너지 소비량 자료의 결측치 추정

<부표> 다중치환법을 이용한 제조업 분야별 에너지 사용량 추정값(계속)

단위: 천 TOE

연도	의료, 정밀, 광학	전기 장비	기타 기계 및 장비	자동차 제조업	기타 운송장비	기타
2000	77.5	354.5	1,038	1,121	204	424
2001	79.0	350.1	1,038	1,111	251	388
2002	82.2	409.9	1,065	1,187	262	363
2003	81.8	398.3	1,035	1,234	270	301
2004	70.8	407.4	1,090	1,369	314	293
2005	77.9	441.1	1,031	1,427	350	243
2006	87.3	440.6	1,018	1,514	384	238
2007	120.7	442.6	985	1,560	378	200
2008	101.4	488.4	1,006	1,547	530	223
2009	105.3	502.4	1,032	1,473	615	191
2010	88.4	377.3	623	1,688	414	294
2011	140.0	527.9	1,066	2,055	563	214
2012	141.9	598.8	988	1,913	591	205
2013	164.5	629.6	1,031	1,942	725	186
2014	166.5	622.5	1,152	1,973	581	239
2015	135.6	520.3	980	2,018	443	243

ABSTRACT

Estimating Missing Data of Industrial Energy Consumption Using a Multiple Imputation Method

Oh-Sang Kwon*, Yun-Sun Park**, Sangjun Lee***

This study estimates 21 industries' missing energy consumption data in Korea using multiple imputation (MI) methods. The pattern of data missing is identified, and the strategies of MI method's addressing item missing data are reviewed. The study chooses each industry's production activity index as the variable to be included in the imputation model. Our study provides a completed data set of industrial energy consumption for period 2000-2015. The results show that industrial production activity index is an appropriate explanatory variable of the imputation model for most energy intensive industries such as energy product industries, chemical industries, metal industries, automobile industries, computer and electronic industries, and etc.

Key Words : Industrial Energy Consumption, Missing Data, Multiple Imputation

* Professor, Department of Agricultural Economics and Rural Development, College of Agricultural and Life Sciences, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-921, Korea (main author). khsang@snu.ac.kr

** Graduate Student, Department of Agricultural Economics and Rural Development, College of Agricultural and Life Sciences, Seoul National University (author)

*** Research Fellow, Korea Energy Economics Institute (author)