공간개념 조사구를 활용한 사업체관리

정 동 욱*

1. 서론

우리가 관심을 갖고 있는 모든 변수를 포함하는 정도 높은 정보를 얻기는 현실적으로 어렵다. 이와같은 현안을 해결하기 위해 여러경로로 수집된 정보를 통해 보강할 필요가 있다. 기존 정보에 없는 변수를 추가하거나 결측치를 보강(imputation)하기 위한 방법으로는 대용량데이터의 병합, 통계적레코드연결, 다원천대체와 같은 데이터 통합 방법이 주로 사용되는데, 이때 자료매칭방법을 이용한다. 자료매칭을 통한 방법은 신규 통계조사를 통해서 데이터를 얻는 것보다 시간과 비용을 절약할 수 있고, 응답자의 부담을 경감시킬 수 있다는 장점을 가지므로 모집단 관리에도 편리성을 제공할 수 있다.

통계청에서는 산업부문별로 전수 또는 표본조사 방식으로 사업체단위 통계조사를 실시하고 있다. 사업체단위 통계조사 결과의 종합적이고 효과적인 관리는 통계조사간 비교분석을 용이하게 하고 조사결과의 정확성을 향상시킬 수 있다. 현재, 중복·누락을 최소화 할 수 있는 사업체 모집단관리시스템을 개발하여 운용 중에 있다.

효율적인 사업체관리를 위한 모집단 구축 및 관리를 경제의 법칙 측면에서 살펴보면, 가장 중요한 부분은 고유번호부여 및 관리와 각종 자료의 matching 기법이다. 본 논문에서는 고유번호부여 및 관리부문과 data matching 부문을 언급하고자 한다. 첫 번째 부문에서는 개체개념(entity concept)과 공간개념(spatial concept)을 적용한 고유번호 접근방법을 다룰 것이며, 두 번째 부문에서는 data matching 기법중 exacting-matching에 대해서 알아보고 다음장에서는 국세청 자료와 같은 행정자료의 실시간 이용이 어려운 경우 효과적인 데이터통합(data fusion) 기법을 검토하여 현재 사업체단위 통계조사의 효율적인 관리를 위한 제언을 하고자 한다.

*통계청 통계개발원

2. 사업체관리를 위한 접근방법

1) 개체개념(Entity Concept)에 의한 접근법

자료레코드 고유ID 부여시 개체를 중심으로 관리하는 방법으로, 최초 생성된 사업체에 고유ID를 부여한 후 사업체가 폐업될 때까지 개체를 유지하는 방법이다. 사람에게 부여된 주민등록번호와 같이 전출입시에도 계속 추적하여 관리해 주는 방법이다. 고유ID부여 및 관리시 사업체에 대한 구속력을 부여하거나 상시 모니터링이 가능하다면 사업체 생소멸통계 작성 등을 위해 가장 효과적인 접근방식이다.

그러나, 국세통계연보에 따르면 2005 말 현재 795천개 업체가 폐업되었고, 880천개 업체가 신규등록 되었다. 전사업체의 약 21.4% 규모로 개체접근법에 의한 고유ID의 유효기간은 5년 이내이며, 전 사업체의 50%정도는 2~3년 이내에 고유ID를 새롭게 부여받는다.

<2005년말 기준 사업체등록현황>

(단위:개,%)

	총계	신규	폐업
합 계	4,121,612	880,716 (21.4%)	795,755 (19.3%)
법인사업자	400,398	66,375 (16.6%)	41,761 (10.4%)
일반과세자	2,117,551	461,862 (21.8%)	393,941 (18.6%)
간이과세자	1,603,663	352,479 (22.0%)	360,053 (22.5%)

또한, 국세자료와 같은 행정자료를 이용하지 못한다면 전출입 사업체를 파악할 때 전국단위로 유사한 사업체를 매칭하여 재확인해야 하는 등 비용과 시간측면에서 효율성의 한계를 갖게 된다.

2) 공간개념(Spatial Concept)에 의한 접근법

사업체가 위치한 장소에 고유ID를 부여하는 방식으로 공간고유화 접근법이다. 공간접근법은 개체접근법과 달리 공간을 고유화하였기 때문에 사업체의 전출입에 관계없이 부여한 고유ID의

변동성이 적게 되며, 장소내의 사업체변동을 data exacting-matching기법에 의해 파악이 가능하여 비용과 시간측면에서 보면 개체접근법보다 사업체관리의 효율성을 갖는다.

즉, data exacting-matching기법에 의해 고유ID가 부여된 공간내의 사업체명, 대표자명, 산업분류 등의 변동과, 더 이상 사업체가 존재하지 않는 공간과 새롭게 사업체가 생긴 공간도 추가비용 투입없이 손쉽게 파악할 수 있다. 사업체의 전출입통계 작성은 개체접근법과 마찬가지로 어렵지만 공간생멸통계는 부수적으로 산출할 수 있다.

공간접근법은 현재 경제활동인구조사와 같은 가구부문 조사에서도 널리 적용되는 고유ID 부여방식으로 조사의 편리성과 공간패널 작성에 유용하게 활용될 수 있다.

3. Data Exacting-Matching기법

데이터보강은 관리하고자 하는 데이터에 기존 정보를 결합하여 양(개체수)과 깊이(변수의 수)를 늘리는 것이다. 이때 정보량을 늘리기 위해 자료매칭기법을 이용하는데 이는 크게 exacting-matching과 statistical matching 기법으로 나눌 수 있다.

Exacting-Matching method는 Donor file이 Recipient file의 모든 개체를 포함하는 경우 사용하는 방법으로, 서로 다른 데이터 파일로부터 matching key를 생성시켜 동일한 개체를 연결하는 방법이다.

Statistical Matching method는 Donor file과 Recipient file의 개체가 서로 상이한 경우, 동일한 개체를 연결하기 보다는 통계적으로 유사한 개체를 연결하는 방법이다.

통계청에서는 매년 실시되는 사업체센서스 자료가 있기 때문에 matching key조합을 통한 Exacting-Matching method를 이용한 data fusion 수행방법을 제안하고자 한다.

통계청에서는 매년 1인이상 전 산업의 사업체를 대상으로 사업체기초통계조사를 실시한다. 이 조사결과를 Recipient file로 정의하며 정보부분(X)과 matching key로 활용될 부분(Y)으로 구분하여 다음과 같은 step으로 file을 구성한다.

step1. 사업체의 위치정보(행정구역분류부호, 조사구번호, 일련번호) 조사구에 근거하여 공간 고유ID를 부여한다.

step2. Recipient 개체 조사시 사업체변동을 파악할 수 있는 조사항목을 추가하여 조사하며, 정의는 다음과 같다.

① 존속: 고유ID가 부여된 공간에 사업체가 있다

② 신규: 고유ID가 부여되지 않았던 공간에 사업체가 있다

③ 폐업: 고유ID가 부여되었던 공간에 사업체가 없다

④ 대상외: 고유ID가 부여되었던 공간의 사업체가 조사대상이 아니다

step3. Donor file을 정의한다. 예를들어 종사자의 상세한 근로형태 data fusion을 할 때, 노동부의 사업체근로실태조사를 Donor file로 정의한다. Donor file도 recipient file과 같은 개념으로 정보부분(Z)와 matching key로 활용될 부분(Y')으로 구분한다.

step4. matching key를 조합한다. Exacting Matching기법에서 matching key는 여러차례 simulation을 통해 one-to-one matching이 가능한 matching key를 찾는다.

step5. macro program으로 data fusion을 수행한다.

(Recipient file)	(X) 조직형태,사업체구분,산업분류 종사자수,매출액	(Y) 행정구역코드,사업체명, 대표자명

+(Donor file)	(Z) 고용형태별종사자수, 교대근무현황, 탄력근무현황 등	(Y) 소재지 ,사업체명, 대표자명

= (Fusion file)	(X)	(Y'')	(Z)

위 step을 통한 data fusion 결과 정보의 깊이가 9개이었던 Recipient file은 Donor file의 세부 근로형태 정보가 추가됨으로써 신규조사 없이 많은 정보를 확장시킬 수 있다.

만약 Donor file의 갯수를 더욱 더 확장한다면 기본항목 9개만 갖고 있는 Recipient file은 기하급수적으로 정보깊이를 늘어날 것이다.

4. 공간변동 통계표 작성

2장에서 언급한 바와 같이 개체접근법에 의한 사업체 생멸은 전출입 사업체의 추적이 곤란하여 현실적으로 작성이 어렵다. 그러나 공간접근법에 의한 고유ID 부여 방식은 공간내의 생멸을 작성할 수 있는데 step은 다음과 같다.

step1. t연도 Recipient file과 t+1연도 Recipient file을 고유ID로 matching 한다.

step2. 사업체변동 부호별로 레코드를 grouping한다.

step3. 존속(Spatial continuance) group을 아래와 같이 4개의 subgroup으로 나눈다.

subgroup1 : 사업체명+대표자명 변경

subgroup2 : 대표자명만 변경

subgroup3 : 사업체명만 변경

subgroup4 : 기타로 구분하며, 4개의 subgroup은 mutually exclusive하다.

step4. 통계표는 다음과 같다.

	존속				신규	폐업	대상외
	sub-group1	sub-group2	sub-group3	sub-group4			
지역별× 산업별× 사업체규모별× 경영조직별 등							

5. 결론

유럽의 Business Register는 국세등록정보와 공유된 시스템으로 기업체의 고유ID를 관리하고 있다. 이 경우 개체접근법에 의한 Recipient File을 관리할 수 있다.

최근 우리나라에서도 통계조사 또는 자료분석시에 행정자료를 활용함으로써 통계자료의 정도를 제고하고, 응답자의 부담을 경감시키고자하는 노력을 해나가고 있다. 그럼에도 불구하고, 사업체의 국세청등록정보와 같은 행정자료의 공유는 현재까지는 이용이 곤란한 상태이다. 사업체의 생멸 또는 이전 등 다양한 정보는 유사사업체를 검색하여 전화를 통한 재질의하는 방법을 제외하고는 현실적으로 얻기가 어렵다고 볼 수 있다. 가구관리에서 사용하고 있는 공간접근법을 사업체 관리에서도 활용하여 사업체의 건물단위까지의 고유ID를 부여한 후 사업체의 생멸 또는 이전을 지속적으로 관리하도록 하는 것을 검토해 보자.

매년 실시하는 사업체기초통계조사를 Recipient File로 구축한 후, 세부 업종별 통계조사 결과를 Donor File로 이용한 Data Fusion으로 정보의 깊이를 늘릴 수 있으며 보다 다양한 분석을 가능하게 할 것이다.

참고문헌

정성석, "데이터보강을 위한 데이터 통합기법에 관한 연구", 2004

이건, "2005년 인구주택총조사의 조사구설정 방법", 2006

Hans-Urich Zaugg, "The Service for Spatial Data in The Swiss Federal Administration"