통계개발원
Statistics Research Institute

매크로 데이터의 비밀보호기법

국제공동연구 출장결과 보고서

2008. 10.

통계개발원

매크로 데이터의 비밀보호기법
국제공동연구 출장결과 보고서

Ⅰ. 출장개요

가. 출장목적

- 통계이용자들의 자료제공 요구시 개인정보의 비밀보호에 대한 중요성이 대두되어 통계자료의 비밀보호 기법의 습득과 활용이 요구되고 있음

- 이에따라 통계개발원에서는 국제공동연구사업의 일환으로 '06년부터 3개년 계획으로 「통계자료 비밀보호 기법」에 관한 국제공동연구를 추진하였음

- '06~'07년 기간중에는 마이크로 데이터(이산형, 연속형)의 비밀보호 기법연구를 수행하였으며, '08년에는 매크로(집계표) 데이터의 비밀보호에 대한 해외전문가와 국제공동연구 수행을 통해,

  · 매크로데이터의 통계적 비밀보호 방법 관련 선진기법의 습득

  · 동 선진기법을 사업체단위 통계조사 결과에 적용함으로써 개인정보 보호 및 통계자료의 확대 제공을 위한 통계정보의 유용성 확보

  · 비밀보호와 관련된 지속적인 연구 및 실무적용을 위함

나. 출장기간

- 2008. 7. 12 ~ 9. 5 (8주간)

다. 출장자 및 출장지역

소 속	직 급	이 름	출장지역
경제통계실	통계사무관	정 동 욱	미 국
연구기획실	통계주무관	김 경 미	(워싱턴 D.C.)

※ 방문기관 : 미국 국립보건통계센터(National Center for Health Statistics)

라. 주요 수행사항

- 매크로데이터의 비밀보호 기법에 대한 기초연구
  · 선행연구 결과에 대한 문헌연구
- 비밀 보호된 매크로 데이터의 파일작성(SAS 프로그램 이용)
  · 대상자료 : 사업체 단위 통계조사
  · 결과자료 : 매크로데이터의 비밀보호기법 적용파일 작성
- 연구보고서 작성 및 연구결과 발표 등

마. 소요예산

- 소요예산 : 29,000천원 내외

바. 기대효과

- 통계청에서 실시하는 각종 통계조사의 집계표 데이터 제공시 응답자의 비밀보호를 위한 효과적인 적용 기대
- 향후 매크로데이터의 비밀보호 기법에 대한 지속적인 공동연구의 발판 마련

Ⅱ. 연구결과

가. 연구목적

○ 자료의 개별정보에 대한 비밀보호강화와 제공범위 확대 증가

- 통계수요측면에서 보면, 다양한 분석을 위한 세부자료와 신규통계 작성에 대한 수요가 증가하고 있음
- 반면, 통계작성기관과 응답자는 개인 및 사업체의 비밀보호 강화에 대한 요구 증대

○ 두 제약점을 해결하는 대안마련 시급

- 각종 정보의 DB화 자료결합 등으로 개별정보 노출위험이 증가하고 있으며, 정보의 원천인 응답자의 비밀보호 준수를 최우선으로 실현하기 위한 대안 마련 모색

○ 제공범위 확대와 응답자의 비밀보호를 동시에 만족시킬 수 있는 비밀보호 기법 마련을 위한 연구 필요

나. 연구내용

○ 대상자료 : 자료제공 형태중 집계표 데이터(매크로 데이터)

○ 집계표 데이터형태

- 집계표는 크게 두 가지 데이터 형태로 구성되어 있음
  - ·빈도데이터(Count Data) : 개수 자체에 의미가 있는 자료
  - ·크기데이터(Magnitude Data) : 개수 보다는 양적인 부분에 의미가 있는 자료
- 한국과 미국의 집계표 형태 비교
  - ·한국 : 표두(Column) 항목에 빈도데이터와 크기데이터 혼합형
  - ·미국 : 빈도데이터와 크기데이터를 각각 구분하여 집계표를 구성

○ 집계표 데이터의 비밀보호 방법

- 노출위험이 있는 민감한 셀 찾는 방법

  · 빈도데이터 : 기준(Base)값 이하인 셀을 노출위험이 있는 셀로 정의

  · 크기데이터

  ① (n,k) dominance rule

  셀을 구성하는 사업체별 조사결과가 셀 합계의 k%이상을 차지하는 경우

  ② p-percent rule

  셀을 구성하는 사업체별 조사결과중 두 번째로 큰 값(X2)이 가장 큰 값(X1)을 p% 이내에서 추정 가능한 경우

  ③ p/q- ambiguity rule

  p% rule을 강화한 방법으로 개별정보중 일부를 알고 있을 가능성이 있을 때 적용하는 방법

  민감한 셀 : $S_{pq}(X) = x_1 - (\frac{q}{p}) \times x_{3+} > 0$

- 집계표 데이터 형태에 따른 비밀보호 방법

빈도데이터	크기데이터
반올림(Rounding) 셀 감추기(Cell suppression) 자료변조(data perturbation) 분류구조변경(data collapsing) 기타(swapping, switching 등)	셀 감추기(Cell suppression)

① 반올림(Rounding)

- 전통적인 반올림 : 기준(Base)를 정하고 0과 기준값(B)으로 변환하는 방법

- 제어된 반올림 : 모든 셀값이 기준(Base)의 배수가 되도록 가감을

반복하는 방법(주변합과 총합을 바꾸지 않으면서 반올림)

- Zero Restriction 50/50 Rounding : 0과 기준값(B)으로 변환하는 과정에서 난수를 발생하여 확률적인 개념에서 반올림을 하는 방법

② 셀감추기(Cell Suppression)

- 노출위험이 있는 셀만 감추는 방법(SODC) : 기준(Base)이하의 셀만 감추기 하는 방법

※ 집계표내의 선형관련식으로 감추어진 셀을 찾을 수 있음

- 보조셀감추기(CCS) : 직접적인 노출이 있는 셀을 감추기 위해 주변의 셀을 보조적으로 감추기 하는 방법

※ 집계표내의 선형관련식으로 감추어진 셀을 찾을 수 없는 장점이 있음

③ 기타

- CTA(Controlled Tabular Adjustment)
- 분류구조방법(Data Collapsing) 등

○ 비밀보호된 집계표 데이터 작성 (별첨보고서 참조)

- 분석대상 자료 : 2006년 기준 광업 및 제조업 통계조사
  중분류 27. 제1차 금속산업

<분석대상의 주요지표>

(단위 : 개, 명, 10억원)

	사업체수	월평균 종사자수	출하액	부가가치	유형자산 연말잔액
제조업 전체	119,181	2,910,935	909,067	326,844	310,176
중분류27	3,173	117,684	88,721	25,282	30,990
구성비	2.7%	4.0%	9.8%	7.7%	10.0%

- 적용기법 : 보조셀 감추기(Complementary Cell Suppression)

목적함수(1): $V(a) = \min \sum a_{ij} z_{ij}$

목적함수(2): $V(a) = \min \sum z_{ij}$

· (1) (2)중 제한되는 정보량을 최소화하는 목적함수(1)을 적용

· 최적경로 탐색은 3가지 접근법중 네트워크기법 적용

ⅰ)결합기법(combinational technique) ⅱ)선형기법(linear programming) ⅲ)수학적 네트워크기법(mathematical network)

- 세부 적용단계

Step 1 : PDC(Primary Disclosure Cell)을 정의함

Step 2 : <표1>과 같이 임의의 PDC(Primary disclosure cell) a_{81}을 출발점으로 가능한 모든 경로(Circuit)를 찾은 후, 각 경로에 해당하는 셀 값의 합을 구함

Step 3 : 경로 중 「0」 값을 포함하고 있는 경우, 이를 찾아 제외시킨다. 왜냐하면 「0」값은 보고서를 통해 이용자에게 제공해야 하는 정보로서 보조감추기 셀로 이용할 수 없기 때문임

Step 4 : Step1, Step2에서 남은 경로 중 감추어진 셀 값이 최소인 경로를 선택함 <표1>

Step 5 : 모든 PDC에 대한 보조감추기 경로를 찾아 총합을 최소화하는 셀을 선택함<표2>

Step 6 : 동일한 방법으로 PDC a_{71}, a_{53}, a_{63}에 대한 최소경로 탐색결과는 <표3>과 같음

〈표1〉 보조감추기 셀 찾기($a_{81}=2$)

PDC	Circuit			감추어진 셀값의 합	찾기
(8,1)=2	(8,2)	(7,2)	(7,1)	13	최소경로
	(8,2)	(6,2)	(6,1)	13	X
	(8,2)	(5,2)	(5,1)	31	
	(8,2)	(4,2)	(4,1)	62	
	(8,2)	(3,2)	(3,1)	154	
	(8,2)	(2,2)	(2,1)	222	
	(8,2)	(1,2)	(1,1)	225	
	(8,3)	(7,3)	(7,1)	4	X
	(8,3)	(6,3)	(6,1)	3	X
	(8,3)	(5,3)	(5,1)	8	X
	(8,3)	(4,3)	(4,1)	16	X
	(8,3)	(3,3)	(3,1)	70	X
	(8,3)	(2,3)	(2,1)	97	X
	(8,3)	(1,3)	(1,1)	96	X

〈표2〉 보조셀감추기(CCS) 결과 (세분류)

산업분류 / 종사자규모	D2721	D2722	D2729
계	217	467	64
5 ~ 9	72	147	22
10 ~ 19	79	137	16
20 ~ 49	47	101	21
50 ~ 99	CSC(11)	45	CSC(3)
100 ~ 199	CSC(4)	21	PDC(1)
200 ~ 299	0	CSC(7)	PDC(1)
300 ~ 499	PDC(2)	CSC(5)	0
500명 이상	PDC(2)	CSC(4)	0

〈표3〉 보조감추기 셀 찾기(a_{71}, a_{53}, a_{63})

PDC	Circuit			감추어진 셀값의 합	찾기
(7,1)=2	(7,2)	(8,2)	(8,1)	13	최소경로
(5,3)=1	(5,1)	(4,1)	(4,3)	19	최소경로
(6,3)=1	(6,2)	(5,2)	(5,3)	30	최소경로

- 타당성 검증

  · 감추어진 셀의 수와 정보량을 통해 노출위험이 있는 셀만 감추기하는 방법과 보조셀 감추기 방법을 비교<표4,5>

  · CCS방법이 SODC 적용결과 보다 감추어진 셀과 정보량이 크나, 이는 SODC방법에 의해 충분한 비밀보호가 이루어지지 않았기 때문임

<표4> CCS와 SODC방법 적용결과 비교 (감추어진 셀의 수)

산업분류		D272	D2721	D2722	D2729
원자료	셀수(A)	24	8	8	8
CCS방법	감추어진 셀 수(B)	10	4	3	3
	C=(B/A)×100	41.7	50.0	37.5	37.5
SODC방법	감추어진 셀수(D)	4	2	0	2
	E=(B/A)×100	16.7	25.0	0.0	25.0
Difference	**C - E**	**25.0**	**25.0**	**37.5**	**12.5**

<표5> CCS와 SODC방법 적용결과 비교 (감추어진 정보량)

산업분류		D272	D2721	D2722	D2729
원자료	정보량(H)	748	217	467	64
CCS방법	감추어진 정보량(I)	40	19	16	5
	J=(I/H)×100	5.4	8.8	3.4	7.8
SODC방법	감추어진 정보량(K)	6	4	0	2
	L=(K/H)×100	0.8	1.8	0.0	3.1
Difference	**J - L**	**4.6**	**7.0**	**3.4**	**4.7**

  · 선형관련식에 의한 재계산 여부를 확인하여 효과적인 비밀보호 여부 체크

$$a_{41}+a_{43}=14$$
$$a_{51}+a_{53}=5$$
$$a_{62}+a_{63}=8$$
$$a_{71}+a_{72}=7$$
$$a_{81}+a_{82}=6$$
$$a_{41}+a_{51}+a_{71}+a_{81}=19$$
$$a_{62}+a_{72}+a_{82}=16$$
$$a_{43}+a_{53}+a_{63}=5$$
$$a_{ij}\geq 1$$

<표2>의 결과를 예시로 위와 같은 선형관련식을 세우고 선형방정식을 풀면 유일한 해를 찾을 수 없기 때문에 성공적인 비밀보호가 이루어진 것으로 검증

※ 자세한 분석내용은 “연구결과보고서” 참조

다. 향후계획

○ 광업 및 제조업 통계조사에 대한 비밀보호 방법의 추가 연구수행

- 집계표 구성형태를 데이터유형(빈도 및 크기데이터)에 따라 구분하는 집계표 설계방법 연구
- 데이터유형에 따른 다양한 민감한 셀 찾는 방법 적용 및 기법적용에 따른 정보제공 효율성에 관한 연구 수행

○ 산업구조통계조사에 대한 비밀보호 방법 확대 적용방안 마련

- 통계청에서 작성중인 산업구조통계조사에 대한 새로운 비밀보호 기법 확대 적용방안 연구

Ⅲ. 별첨

□ 연구결과보고서

「매크로데이터의 비밀보호방법에 대한 고찰」