생활인구 산정을 위한 가명정보의 결합키연계정보 생성방안

이수미, 이정숙*, 김은진**, 신용태***

숭실대학교

leesumi@soongsil.ac.kr, *jungsuk2023@soongsil.ac.kr, **sonaflux30@soongsil.ac.kr,
***shin@soongsil.ac.kr

A Study on the Generation Method of Combination Key Linkage Information of Pseudonym Information for Estimating De Facto Population

Sumi Lee, Jungsuk Lee*, Eunjin Kim**, Yongtae Shin***

Soongsil Univ.

요 약

이 연구의 목적은 생활인구 산정을 위한 대용량 데이터 연산과 결합키연계정보 생성의 업무 부담을 감소시키고, 기반 데이터의 정확성을 높이는 방안을 제안한다. 생활인구는 주민등록인구, 체류인구, 외국인인구를 포함하며, 가명정보 결합을 통해 월별로 측정된다. 연구는 데이터의 효율적인 결합을 위해 시계열 키를 제외한 결합키 구성을 제안하고, 중복 결합키 문제에 대응하기 위한 대안법을 제시한다. 이 연구를 통해 향후 가명결합 관련 정책 수립 및 실행에 있어 중요한 기초 자료로 활용될 수 있을 것으로 기대한다.

Ⅰ. 서 론

저출산, 고령화와 더불어 일자리 부족으로 인해 지방 중소도시의 인구 감소가 지속되고 있다. 한국고용정보원의 자료에 따르면, 전국 228개 시군구 중에서 소멸 위험이 있는 지역은 2013년 75곳에서 2022년에는 113곳으로 늘어나고 있다. 행정안전부는 2021년 89개 시군을 인구감소지역으로 지정하고, 지역 소멸에 대응하기 위해 2022년부터 매년 1조 원 규모의 지방소멸대응기금을 이들 지역에 투자하고 있다. 또한, 2022년 '인구감소 지역 지원 특별법'에 '생활인구' 개념을 도입하여 기존의 정주인구 중심의 인구 관리 정책을 체류인구를 포함하는 개념으로 확대하고 있다.[1] 생활인구는 거주가 아닌 생활을 중심으로 인구를 바라보는 새로운 모델로, 직장, 학교, 관광, 휴양 등을 목적으로 체류하는 인구를 포함한다.[2] 이는 지역의 활력을 높이는 사람들까지 포함하여 국가 총인구 감소 상황에서 보다 현실적인 방안이며, 교통과 통신 발달로 인한 이동성 증가, 여가 중시, 일과 생활의 균형 등 현대 트렌드를 반영한다.[3]

본 연구에서는 인구감소 지역의 지원을 위해 신설된 생활인구 개념을 기반으로 명확한 선정기준과 측정을 통해 생활인구의 합리적인 활용방안을 모색하고자 한다. 생활인구 측정을 위한 활용 데이터는 행정안전부주민등록, 법무부외국인등록, 재외동포거소신고 자료와 통신3사(SK텔레콤, KT, LG U+)의 모바일 이동 자료를 가명결합한 데이터이다.[4] 가명정보 결합을 수행하려면, 결합키관리기관이 결합키연계정보를 생성하여 결합전문기관에 전송해야 하며, 결합전문기관은 이 정보를 바탕으로 가명정보 결합을 진행한다. 선행 연구들은 주로 인구 감소와 지역 활성화에 초점을 맞추어 생활인구를 연구해 왔으나, 가명정보 결합에 있어 결합키연계정보의 생성과 관련된 연구는 아직 미흡하다.

따라서 본 연구는 주민정보와 이동통신정보의 결합을 위해 필요한 결합키연계정보를 효율적으로 생성하기 위해 다음과 같은 방안을 제시한다.

1. 결합키연계정보 생성 시 소요되는 리소스 및 시간을 줄이는 방법을 검토한다.

2. 결합키 중복으로 인한 데이터 누락을 줄이는 방안을 제시해보고자 한다.

Ⅱ. 관련 연구

1. 생활인구

생활인구는 2023년부터 도입된 개념으로, 현대 사회의 이동성 증가를 반영하며 주민등록인구, 일시 체류인구, 그리고 외국인을 포함한다. 「인구감소지역 지원 특별법」에 따라 생활인구는 주민등록인구, 특정 목적으로 월 1회 이상 특정 지역에 3시간 이상 머무는 체류인구, 그리고 외국인 등록을 한 인구로 구분되고 [그림 1]과 같다.[5]

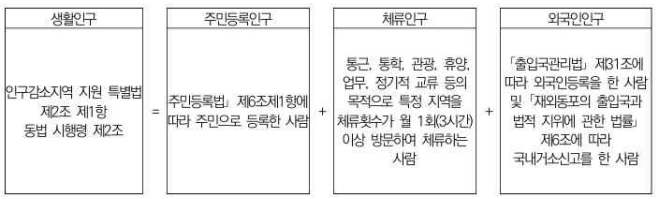

그림 1. 「인구감소지역 지원 특별법」상 '생활인구'의 개념[1]

Fig 1.Concept of 'Living Population' under the Special Act for Support in Population Decline Areas[1]

이 외에도 체류인구의 정의는 다양한 해석이 있다. 안소연 외(2023)는 주민등록이 되어 있지 않은 상태에서 특정 지역에 1박 이상 머무르는 인구로 정의하여 법률의 정의와 차이가 있다.[6] 이 외에도 통계청(2020)은 유동인구를 한 지역에서 2시간 이상 체류한 후 다른 지역으로 이동하여 그곳에서도 2시간 이상 머무르는 경우로 정의한다.[7] 이같이 체류인구는 각 지자체의 체류 시간과 빈도에 따라 측정되며 다양한 해석이 가능하나 일반적으로 3시간 이상 머무르거나 1박 이상 체류하는 인구를 포함한다. 이런 정의는 지역의 인구 활동성 파악에 중요한 기준이다.

2. 가명정보 결합

가명정보 결합이란 서로 다른 개인정보 처리자들이 각자 관리하는 정보를 가명 처리한 후 결합할 수 있다. 개인정보보호위원회이나 관련 중앙행정기관의 장이 지정한 결합전문기관이 결합 작업을 수행한다(보호법 제28조의3 제1항). 가명정보 결합은 ① 결합신청자의 결합신청, ② 결합키관리

기관의 결합키연계정보 생성, ③ 결합전문기관의 가명정보 결합 및 반출, ④ 결합신청자의 반출정보 활용 및 관리 등으로 진행된다.[8] 가명결합 프로세스를 결합키연계정보의 생성 중심으로 표현하면 [그림 2]와 같다.

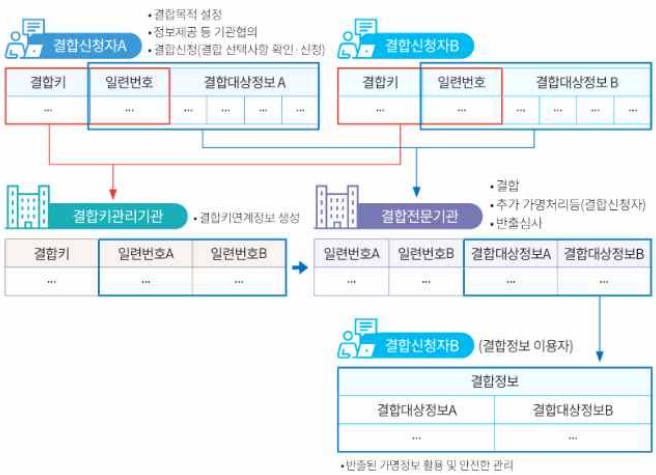

그림 2 결합키연계정보 생성 및 결합 프로세스[8]
Fig 2. Linkage Key Information Creation and Combination Process[8]

III. 연구 문제

1. 결합키연계정보 생성 시 리소스 및 시간 소모

생활인구 산정을 위한 가명정보 결합 과정에서 결합키관리기관은 결합키연계정보를 생성해야 한다. 결합키(64byte), 일련번호(20byte)로 정의하면, 주민정보와 외국인정보의 데이터는 약 12GB(1억 5천만건, 3개월), 통신사 데이터는 약 144GB(18억건, 3개월)가 된다. [그림 3]와 같이결합의뢰기관의 결합키/일련번호의 크기가 커서 한 번에 결합키연계정보를 생성하기 어려워진다. 따라서 결합키/일련번호 쌍을 결합키관리기관에 분할하여 전송하고, 결합키관리기관은 이를 받아 각각의 결합키연계정보를 생성한 후 이를 결합전문기관에 전달한다.

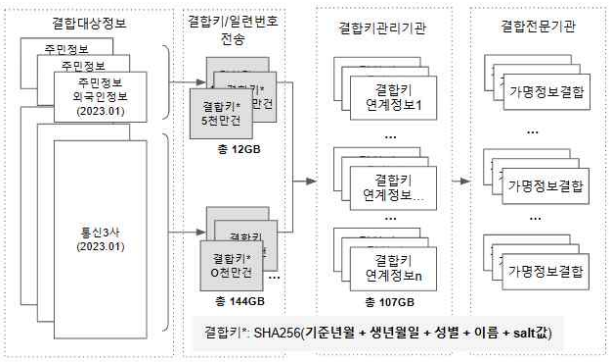

그림 3. 분할 생성된 결합키연계정보를 이용한 가명정보결합
Fig 3. Pseudonymized Data Combination Using Segmented Generated Linkage Key Information

이러한 반복적인 프로세스는 상당한 시간과 리소스가 소모된다. 데이터 결합의 복잡성과 리소스 소모를 줄이기 위해 결합키연계정보 생성 방법의 개선이 필요하다.

2. 중복 결합키 삭제로 인한 생활인구 데이터 누락

데이터 결합 과정에서 결합키는 두 데이터셋을 연계하기 위한 고유한 식별자로 사용된다. 그러나 입력되는 데이터가 같으면 해시 결과도 동일하기 때문에 중복이 발생할 수 있다. NICE신용평가정보에 따르면, 주민등록상 생년월일이 일치하는 사람 중 같은 이름을 가진 경우가 11.7%라는 보고가 있다.[9] 중복 결합키를 포함하여 결합하게 되면 데이터가 의도와는 다르게 해석될 수 있으므로 일반적으로 삭제한다. 생활인구 결합의 경우 중복 결합키를 삭제하면 약 500만 명의 정보가 누락 된다. 정확한 인구 데이터를 확보하기 위해서는 이러한 문제에 대한 보완이 필요하다.

IV. 개선 방안

1. 결합키연계정보를 위한 대상데이터 사이즈 축소

효율적인 결합키연계정보 생성을 위한 개선방안으로, 결합의뢰기관은 중복을 제거하고 유일한 결합키/일련번호 쌍으로만 구성된 데이터를 결합키관리기관에 전송한다. 즉, [표 1]과 같이 결합키를 생성할 때 기준년월을 입력값에서 제외한다.

표 1. 결합키 구성 방법 제안
Table 1. Proposed Method for Linkage Key Configuration

기존	SHA256(기준년월+**생년월일+성별+이름**+salt)
개선	SHA256(**생년월일+성별+이름**+salt)

이 과정에서 주민등록정보와 외국인정보의 데이터 크기는 [그림 4]과 같이 약 1/3(약 5천만건), 통신사 데이터는 약 1/30(약 6천만건)로 줄어든다.

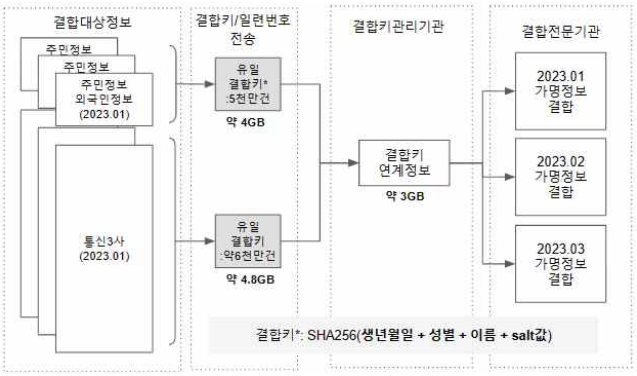

그림 4 개선된 결합키연계정보 생성 방안
Fig 4. Proposed Method for Improved Linkage Key Information Generation

이를 통해 결합키연계정보의 크기가 현저히 감소하고, 데이터의 관리가 용이해진다. 즉, 결합키관리기관은 한 번의 데이터 처리로 필요한 마스터 역할을 하는 결합키연계정보를 생성할 수 있고, 시간과 리소스를 절약할 수 있다.

2. 중복 결합키를 결합키연계정보에 추가

결합키 중복으로 인한 데이터 손실을 해결하기 위한 절차로, ① 중복 결합키가 발생하는 정보를 분리한다. ② 유일한 결합키가 있는 주민정보를 기준으로 체류정보(통신)를 join하여 결합키연계정보 생성한다. ③④ 분리했던 중복 결합키의 일련번호들을 결합키 연계정보에 추가한다. 이는 [그림 5]와 같다.

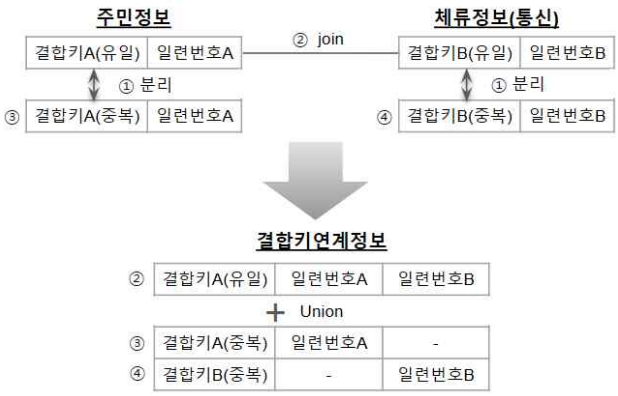

그림 5. 결합키연계정보에 중복 결합키의 일련번호 정보 추가
Fig 5. Adding Serial Number Information of Duplicate Linkage Keys to Linkage Key Information

위와 같이 생성된 결합키연계정보를 기반으로 결합을 하게 되면 주민정보와 체류정보에서 결합키를 생성할 때 중복되는 정보를 누락하지 않고 결합을 할 수 있어 데이터의 손실을 줄일 수 있다.

V. 결 론

본 연구는 생활인구 산정 과정에서 가명정보 결합 효율성 저하 및 중복 결합키로 인한 데이터 손실 문제를 해결하기 위한 방법을 제시하였다.

첫째, 결합키 생성 시 시계열 정보를 제외하여 키를 구성하면 결합키연계정보의 크기와 생성에 필요한 자원을 줄일 수 있다. 현재 가명정보 처리 가이드라인에는 결합키연계정보가 결합키/일련번호 매핑을 기준으로 생성되는데, 결합키에 시계열 정보를 포함하는 경우 데이터 양이 증가하면서 결합키연계정보의 크기와 생성에 필요한 자원이 증가하게 된다. 시계열 정보를 다루는 결합키의 경우 결합키연계정보를 한 번 생성하고 반복적으로 활용할 수 있도록 함으로써, 가명결합 과정의 효율성을 개선할 수 있다. 이러한 내용이 향후 가이드라인에 포함된다면 가명정보 결합이 필요한 연구자들에게 유용한 참고자료가 될 것이다.

둘째, 결합키 중복 문제를 해결하기 위한 방법은 주민등록번호를 결합키로 사용하는 것이 가장 효과적이지만, 이는 현재 법적 제한이 있어 결합키를 생성하는데 어려움이 있다. 이에 대안적 방법으로 중복 결합키를 결합키연계정보에 포함시키는 방법을 제안하고, 이러한 접근은 데이터의 품질 개선 및 생활인구 산정의 정확성과 신뢰성을 높이는데 기여할 것으로 생각된다.

끝으로 본 연구에서 제안하는 해결 방안이 가명정보 결합에 관한 정책을 수립하고 실행하는 데 있어 중요한 자료로 활용되길 기대한다.

참 고 문 헌

[1] S. S. Lee, B. H. Yun, M. H. Lee and Y. H. Kwon, “Consideration of issues for calculating de facto population and utilizing urban policies,” *Journal of Urban Studies and Real Estate*, 14(4), pp. 69-87, 2023. (https://doi.org/10.21447/jusre.2023.14.4.4)

[2] C. W. Seo and J. A. Bae, “What are the Drivers Affecting the Formation of the Living Population? : Focusing on the Effects of the Local Culture and Tourism Financial Expenditure,” *The Korea Local Administration Review(Krila)*, 37(4), pp. 222-240, 2023. (http://dx.doi.org/10.22783/krila.2023.37.4.221)

[3] B. K. Min and J. S. Choi, “Characterizing regions based on the de facto population: Focusing on the local areas in South Korea,” *Journal of the Korean Urban Management Association*, 36(4), pp. 41-60, 2023. (https://doi.org/10.36700/KRUMA.2023.12.36.4.41)

[4] Statistics Korea and Ministry of the Interior and Safety, “Results of the De Facto Population Estimation in Pilot Areas Based on Pseudonymized Data Combination between Public and Private Sectors,” Jan. 2024.

[5] Legislation and Law Information Center, “Special Act for Support in Population Decline Areas,” 2023.

[6] S. H. Ahn, S. J. Lee, S. H. Min, M. A. Kim, B. K. Jeon and M. S. Kang, “The Necessity of Introducing Transient Population in the Era of Population Decline and Policy Measures,” *National Land Policy Brief*, pp. 1-8, 2023.

[7] Statistics Korea, “Mobile Telecommunications Big Data-based Floating Population Mapping Service,” 2020, Retrieved May 12, 2024, from https://giraf.sktelecom.com/web/kostat/.

[8] Personal Information Protection Commission, “Guidelines for the Processing of Pseudonymized Information,” April. 2022.

[9] Yonhap News Agency, “According to NICE Credit Information, there is an 11.7% probability that individuals with the same date of birth and name exist in the resident registration records,” Sep. 17, 2012, Retrieved May 15, 2024, from https://www.yna.co.kr/view/AKR20120914199900002