The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 849–861

DOI: https://doi.org/10.5351/KJAS.2021.34.5.849
Print ISSN1225-066X / Online ISSN 2383-5818

A study on the probabilistic record linkage and its application

Yeonok Choi[a], Sangin Lee [1, b]

[a]Statistics Korea, and Department of Information and Statistics, Chungnam National University;
[b]Department of Information and Statistics, Chungnam National University

Abstract

This paper aims to introduce the basic concept of probabilistic record linkage and its statistical framework, and describe the specific process and principle of performing it using a real example from Statistics Korea. First, we briefly describe the deterministic record linkage and compare it with probabilistic record linkage. We introduce the Fellegi-Sunter model framework for record linkage and the related paprameters: m-probability, u-probability, matched weight and decision rule. Finally, we show the detailed process of record linkage under Fellegi-Sunter model framework and evaluate the record linkage results, using sample data from the registered-based census and Population and Housing Census survey in Statistics Korea.

Keywords: probabilistic record linkage, Fellegi-Sunter model, m-probability, u-probabililty, match weight

1. 서론

빅데이터 활용이 중요해지면서, 통계정보의 생산에 있어서도 조사자료 뿐만 아니라 행정자료 등 다양한 출처(source)의 자료들을 활용하는 것이 크게 늘어나고 있다. 여러 출처의 데이터를 연계(linkage)하거나 결합(integration)함으로써 응답 부담이나 비용은 절감하고 통계자료의 종류와 내용 및 활용성은 증대시킬 수 있다. 해외에서는 수십여 년 동안 자료 연계(record linkage)에 관한 연구가 발전해 왔으며, 정부와 민간에서 매우 광범위하게 활용되고 있다. 특히 공공분야에서 자료연계 방법론을 적용한 연구가 활발한데, 센서스 자료를 다른 표본조사 또는 행정자료와 연계하거나, 보건 분야에서 인구 자료와 질병, 사망, 진료기록 등을 연계한 사례들이 많이 소개되고 있다. 우리나라의 경우 상대적으로 자료의 접근성에 대한 제약이나 활용 수요의 부족 등으로 방법론에 대한 연구가 별로 이루어지지 못하였다. 그러나 우리도 최근 데이터 연계의 중요성이 강조되고 있으며 단일 자료만으로는 복잡한 사회 문제에 대응하는 데 한계가 있기 때문에 학술적 · 정책적 연구에 있어 데이터 연계는 점점 더 주목받게 될 것이다.

자료를 담고 있는 각각의 데이터베이스는 어떤 기관의 특정한 목적을 위해 구축되어 있다. 이러한 각기 다른 원천의 자료를 결합한다고 할 때, 크게 두 종류의 접근 방식이 있다. 하나는, 결합시키는 대상(개인 또는 사업체)이 동일 개체는 아니지만 같은 특성을 공유하는 유사한 개체를 찾아 해당 레코드를 활용하는 것이다. 다른 하나는, 동일 개체를 대상으로 공통의 식별변수(identifier)를 매개로 레코드끼리 연계하여 결합시키는 것이다. 전자를 통계적 매칭(statistical matching)이라 하며, 후자를 정확 매칭(exact matching) 또는 자료 연계(record linkage)라고 부른다. 다시 말해서 자료연계란 여러 데이터 파일에 존재하는 동일한 개체(개인, 가구,

This paper was supported by the National Research Foundation of Korea (NRF) grant (No. 2020R1I1A3071646).
[1] Corresponding author: Department of Information and Statistics, Chungnam National University, 99, Daehak-ro, Yuseong-gu, Daejeon 34134, Korea. E-mail: sanginlee44@gmail.com

Published 31 October 2021 / journal homepage: http://kjas.or.kr

사업체 등)에 대한 다양한 정보를 하나의 데이터로 결합시키는 과정을 의미한다. 한 데이터베이스 내에서 중복된 기록을 찾아내거나 다수 데이터베이스의 자료를 연계하는 것을 모두 포함하는 개념이다.

주민등록번호와 같은 고유식별번호(unique identification number)를 사용할 수 있다면 다른 데이터 파일로부터 동일 인물을 쉽게 확인할 수 있고 해당 개인의 자료를 연계시키는 것은 매우 쉬운 작업이다. 그러나 고유번호를 사용할 수 없을 때 주로 이름, 생년월일, 주소 등 간접 변수들을 활용하여 동일인인지를 판단함으로써 자료 연계를 하게 되는데, 통상 수집된 자료에 중복, 누락, 오타 등 여러 오류가 내재되어 있기 때문에 연계 과정이 그리 간단하지 않다. 따라서 자료 연계의 관건은 여러 다른 원천의 데이터 파일에서 어떻게 동일 개체를 식별해 내느냐에 달려 있다.

자료연계에 관한 연구는 Dunn (1946)이 공공보건 연구에서 개인단위 자료 연계의 필요성을 주장하면서 발전하기 시작하였다. Dunn (1946)은 현대적 자료 연계 이론의 확률론적 기초 아이디어를 제시하였고, Fellegi와 Sunter (1969)가 그들의 선구적 논문 'A theory for record linkage'에서 Newcombe의 아이디어를 수학적 모델로 공식화한 후, Fellegi-Sunter 이론은 지금까지 자료 연계 응용연구의 토대를 형성하고 있다. Fellegi와 Sunter (1969)의 논문 이후 1970년대부터 90년대 사이에 확률적 자료연계 방법론 연구와 실제 데이터 적용이 활발하게 이루어졌다. 주요 연구주제는 최적의 비교함수 찾기 (Jaro, 1989; Winkler, 1990), 모수추정 (Winkler, 1990; Winkler 등, 1993; Winkler, 1995), 정확성 평가 (Hand와 Christen, 2018; Christen과 Goiser, 2007) 등에 관한 것이었다. 2000년대 들어서는 컴퓨터 과학분야에서 머신러닝의 분류 알고리즘을 활용한 자료연계가 발전하였다 (Christen, 2007; Elfeky 등, 2003; Feigenbaum, 2016; Goeken 등, 2011).

본 논문의 목적은 확률적 자료연계 방법의 기초개념과 방법론, 연계과정에 대한 보다 상세한 설명을 통해 독자의 이해를 돕는 동시에, 동 방법론의 실제 활용 가능성을 가늠해 보는데 의의가 있다. 논문의 구성은 2장에서 자료연계 방법으로서 결정적 자료연계(deterministic record linkage)와 확률적 자료연계의 개념적 차이를 설명하고, 3장에서는 Fellegi-Sunter 모형을 기반으로 한 확률적 자료연계에서 다루는 비교쌍의 생성, m확률과 u확률 추정, 가중치, 연계 결정기준 등 대하여 설명한다. 4장에서는 통계청의 등록센서스와 인구총조사 자료로부터의 시뮬레이션용 데이터를 통해 확률적 자료 연계가 수행되는 과정과 수행 결과를 보여준다.

2. 자료연계방법의 종류

2.1. 결정적 자료 연계(deterministic record linkage)

결정적 자료연계는 두 개의 레코드를 비교하여 이름, 생년월일, 주소 등과 같은 연계 변수들의 값이 정확하게 일치할 경우 매치(match) 즉, 동일한 개인으로 판정하고, 일부라도 일치하지 않을 경우 비매치(non-match)로 분류한다.

결정적 연계에서는 일반적으로 연계변수의 중요도는 동일하다. 예를 들어, 두 개인의 레코드를 이름, 생년월일, 주소를 사용하여 비교한다고 할 때, 이름의 각 문자가 일치하고, 생년월일의 숫자가 일치하고, 주소의 각 문자가 완전히 일치할 때만 두 레코드는 같은 사람으로 연계될 수 있다. 만약, 주소 같은 연계변수에 오타가 있거나 표기가 살짝 달라서 조금이라도 일치하지 않는다면 불일치로 판정될 것이다. 따라서, 결정적 자료연계는 실제 매치인데 비매치로 판정받는 개체가 많이 발생할 가능성이 높다. 예로서 Table 1을 가정해 보자. 파일 A와 파일 B에 각 3명의 개인에 대한 레코드가 있다. 두 파일이 실제로는 같은 사람에 대한 자료이지만 두 파일을 결정적 자료 연계에 의해 매치시켰을 때, 결과는 모든 연계변수가 완벽하게 일치하는 1개의 비교 쌍만 매치로 판정된다(Table 2). 결정적 자료연계에서는 연계변수들이 부분적으로 일치하는지는 고려되지 않는다. 한편, 실제로는 동일인이 아닌데 우연히 모든 연계변수가 일치하는 경우가 발생할 수 있다. 이런 경우에도 결정적 자료연계에서는 일단 매치로 처리된다. 다만, 실제 데이터에서는 일치하는 개인이 또 있을 가능성이 매우 크다. 이 때는 다른 추가 변수들(가구원, 혼인년도 등)을 사용하여 진짜 동일인을 찾아내어 연계해야 할 것이다.

Table 1: Deterministic record linkage (example)

파일 A					파일 B				
순번	이름	성별	생년	도로명	순번	이름	성별	생년	도로명
1	홍길동	남	1987	다산로 34길	1	홍길동	남	1987	다산로
2	박철수	남	1995	천호대로 85길	2	박철순	남	1995	천호대로 85길
3	김영희	여	1970	삼양로 27길	3	김영희	여	1970	삼양로 27길

Table 2: Decision by deterministic linkage

이름1	이름2	이름	성별	생년	도로명	판정결과
홍길동	홍길동	1	1	1	0	Non-Match
박철수	박철순	0	1	1	1	Non-Match
김영희	김영희	1	1	1	1	Match

2.2. 확률적 자료연계(probabilistic record linkage)

확률적 자료연계는 잠재적 레코드쌍에 대해 두 레코드의 연계변수가 일치하는 정도에 따라 연계변수별로 점수를 매겨서 전체 점수가 높은 쌍을 동일 개체로 판단하는 방법이다. 데이터 수집과정에서 발생하는 오타나 입력오류, 입력형태 등으로 인해 동일개체라 하더라도 레코드가 완전히 일치하기는 어렵기 때문에 일치 정도를 점수화하여 확률적으로 접근하는 것이다. 확률적 자료연계 방법론은 Fellegi와 Sunter (1969)가 제시한 모형을 토대로 발전되어 왔다.

3. Fellegi-Sunter 자료연계 모형

3.1. 기본모형식

Fellegi와 Sunter (1969)는 확률적 자료연계의 기본 개념을 수학적 모델로 제시하였다. Fellegi-Sunter 모형 하에서 연계하고자 하는 두 개의 파일은 두 모집단을 대표하며, 두 파일 간에 공통요소들이 존재한다는 것을 가정한다. 두 파일 간의 자료연계 뿐만 아니라, 한 파일 내의 중복 레코드를 찾아내는 데에도 적용된다. 기본개념은 두 파일(A, B)로부터 비교(comparison) 가능한 레코드의 모든 쌍들(pairs)에 대해 매치 확률에 따른 가중치를 부여하고, 그 값을 바탕으로 연계(link)-비연계(non-link)-연계 가능성(possible link)을 판정한다.

두 데이터 파일 A, B를 연계한다고 하자. A의 레코드와 B의 레코드를 순차적으로 비교해 나갈 때 생성되는 레코드 쌍들(pairs)의 집합을 비교공간(comparison space, $A \times B$)이라고 하며, 이 비교공간(comparison space)은 이론적으로 참 매치(true match, M)와 참 비매치(true non-match, U) 두 부분집합으로 나누어진다. 어떤 쌍(i, j)이 실제로 같은 사람(또는 개체)으로부터 얻어진 레코드이면 M에 속하고, 다른 사람으로부터 나온 레코드이면 U에 속할 것이다. Fellegi-Sunter는 비교가 이루어지는 레코드 쌍(i, j)이 연계될 수 있는지를 판단하는 평가척도로 다음과 같은 두 조건부 확률의 비율을 제시하였다.

$$R = \frac{P(\gamma \in \Gamma \mid r \in M)}{P(\gamma \in \Gamma \mid r \in U)} \tag{3.1}$$

여기에서, r은 비교쌍, γ는 비교공간 Γ에서 임의의 일치 패턴을 나타낸다. 식 (3.1)의 분자는 어떤 레코드 짝(r)이 실제로 동일 개체(true match)이면서 비교 패턴이 일치(agree)하는 확률이다. 분모는 실제로는 동일 개체가 아닌데(true non-match) 우연히 비교 패턴이 일치(agree)하는 확률이다. 전자를 m-확률(m-probability), 후자를 u-확률(u-probability)이라 부른다.

예를 들어, $A \times B$로부터 선택된 레코드 짝(record pair)을 3개의 연계 변수(예: 이름, 성별, 도로명)에 의해

비교할 때 가능한 일치 패턴(agreement pattern)들은 모두 총 8개(= 2^3)이다. 이때 세 필드가 모두 일치한다면 R은 다음과 같은 형태가 되며,

$$R = \frac{P(\text{agree on name, agree on gender, agree on street name} \,|\, r \in M)}{P(\text{agree on name, agree on gender, agree on street name,} \,|\, r \in U)}$$

이름과 도로명은 일치하고 성별이 불일치한다면 다음과 같이 될 것이다.

$$R = \frac{P(\text{agree on name, disagree on gender, agree on street name} \,|\, r \in M)}{P(\text{agree on name, disagree on gender, agree on street name} \,|\, r \in U)}$$

이 때, 연계변수들이 서로 조건부 독립이라면,

$$\begin{aligned} &P(\text{agree on name, agree on gender, agree on street name} \,|\, r \in M) \\ &= P(\text{agree on name,} \,|\, M) \times P(\text{agree on gender} \,|\, M) \times P(\text{agree on street name} \,|\, M) \end{aligned}$$

로 변환될 수 있다 (Herzog 등, 2007).

Fellegi-Sunter 모형에서 이론적으로 연계변수들간의 독립성 가정이 반드시 필요한 것은 아니다. 그러나 모형의 실제 적용에 있어 모수추정을 단순하고 간편하게 하는 방편으로 독립성 가정을 Fellegi-Sunter는 제시하고 있다. 독립성 가정은 얼핏 매우 강한 가정으로 보이긴 하지만 실제 적용상 크게 중요하지 않을 수 있다. 상호 연관성이 매우 큰 변수들을 연계변수로 사용하지만 않으면 추정값이 크게 왜곡되지 않을 것이다. 예를 들어 주소와 우편번호의 경우 연관성이 높지만, 이름이나 출생일 등은 주소와 별로 관계가 없다. Fellegi-Sunter 모형을 적용한 다른 연구들에서 독립성 가정이 충족되지 않더라도 연계결과가 상당히 정확하다는 연구 결과가 이미 나와 있다 (Winkler 등, 1993). 한편, 최근 들어 머신러닝의 지도학습방법을 사용한 자료 연계 방법론에 관한 연구도 진행되고 있는데 이 방법에서는 독립성 가정을 필요로 하지 않는다.

3.2. m-확률, u-확률 추정

앞에서 기술한 이 모델을 적용하기 위해서는 m-확률과 u-확률을 추정하여야 한다. m-확률은 연계 변수로 사용되는 데이터의 품질과 관련되어 있다. 예를 들어, 성별(gender) 필드를 연계시키는데 비교 짝에서 5%의 오타율이 있다면 m-확률은 0.95가된다. u-확률은 두 레코드가 실제로는 매치가 아닌데 우연히 일치하는 확률이다. 예를 들면, 성별이 우연히 일치할 확률은 1/2이며, 태어난 월이 우연히 일치할 확률은 1/12가 될 것이다. 이론과 달리 현실에서는 참매치 상태를 알 수 없으므로, 이전의 자료연계 경험에서 얻은 사전적인 정보를 이용하거나 기대값-최대화(Expectation-maximization, EM) 알고리즘을 이용하여 m-확률과 u-확률을 추정한다. EM 알고리즘을 사용하면 수치적으로 안정적인 추정 값을 얻을 수 있다. Jaro (1989)는 1995 센서스 자료의 연계를 위한 m과 u-확률 추정에 EM을 적용하였고, Winkler와 Thibaudeau (1991)도 센서스 자료 등의 연계에 EM 알고리즘을 사용하였다.

모든 레코드쌍은 다음과 같이 정의된다고 하자.

$$g_{ij} = \begin{cases} 1 & \text{if record pair } (i, j) \in M, \\ 0 & \text{if record pair } (i, j) \in U. \end{cases} \tag{3.2}$$

이때, 완전한 자료(complete data) 벡터는 $G = \left(\gamma_{ij}, g_{ij}\right)$와 같이 정의되며, 모든 레코드 짝$(i, j)$에 대한 완전한 자료에 대한 우도함수(likelihood function)는 다음과 같은 식으로 표현할 수 있다.

$$f(G \,|\, m, u, p) = \prod_{i<j} pP(\gamma_{ij} \,|\, M)^{g_{ij}} \cdot (1-p)P(\gamma_{ij} \,|\, U)^{1-g_{ij}}. \tag{3.3}$$

여기에서, p는 M에 속하는 레코드 쌍의 비율, $(1-p)$는 U에 속하는 쌍의 비율, $\gamma_{ij} = (\gamma_{ij}^{(1)}, \ldots, \gamma_{ij}^{(k)})^T$는 레코드 쌍 (i, j)의 k개 연계변수들의 일치패턴을 나타낸다. 또한,

$$P(\gamma_{ij}|M) = \prod_{s=1}^{k} m_s^{\gamma_{ij}^{(s)}} (1-m_s)^{1-\gamma_{ij}^{(s)}}, \quad P(\gamma_{ij}|U) = \prod_{s=1}^{k} u_s^{\gamma_{ij}^{(s)}} (1-u_s)^{1-\gamma_{ij}^{(s)}} \tag{3.4}$$

이고, $m_s = P(\gamma^{(s)} = 1|M)$와 $u_s = P(\gamma^{(s)} = 1|U)$는 s번째 연계변수의 m-확률과 u-확률을 나타낸다. 실제, 식 (3.2)의 g_{ij}를 모르기 때문에 우도함수 (3.3)을 직접 최적화 할 수 없고, EM 알고리즘을 적용하여 m과 u-확률을 추정할 수 있다. Jaro (1989)는 EM 알고리즘을 사용하여 g, m, u, p를 얻는 방법에 대해 세세하게 제공하였으며, m 추정치의 초기 값이 u 추정치보다 높기만 하면 EM 알고리즘은 매우 안정적인 모수추정 결과를 제공한다는 점을 논증하였다.

3.3. 매치 가중치(match weight) 산출

산출된 m-확률과 u-확률을 이용하여 각 레코드 쌍 (i, j)에 대한 매치 가중치(match weight; R)를 계산할 수 있다. 앞에서 설명한 대로 R은 m-확률과 u-확률의 비율로 정의되는데, 계산상 편의를 위해 로그를 사용한다.

$$w_{ij}^{(s)} = \begin{cases} \log_2\left(\frac{m_s}{u_s}\right) & \text{if } \gamma_{ij}^{(s)} = 1, \\ \log_2\left(\frac{1-m_s}{1-u_s}\right) & \text{if } \gamma_{ij}^{(s)} = 0. \end{cases} \tag{3.5}$$

연계 필드가 서로 조건부 독립이라고 가정하고 로그를 사용하면, 비교 쌍에 대한 종합가중치는 각 연계 필드들이 갖는 가중치들의 합이다.

$$W_{ij} = \sum_{s=1}^{k} w_{ij}^{(s)} = \sum_{s:\text{일치}} \log_2\left(\frac{m_s}{u_s}\right) + \sum_{s:\text{불일치}} \log_2\left(\frac{1-m_s}{1-u_s}\right). \tag{3.6}$$

3.4. 결정 원칙(decision rule)

종합가중치를 이용하여 레코드 쌍들이 연계(link), 비연계(non-link), 또는 잠재적 연계(possible link)인지를 판정한다. 일정한 컷오프값(cut-off value) 또는 임계값(threshold)을 정하여 이 값보다 큰 종합가중치를 갖는 레코드 쌍을 연계(link)로 분류한다. 임계값이 높을수록 연계되는 쌍은 적어질 것이다. Fellegi와 Sunter (1969)는 두 가지 임계값을 정하여 잠재적 연계와 비연계를 구분할 것을 제안하였는데, 두 임계값은 거짓 연계(false links)와 거짓 비연계(false non-links)의 비율을 고려하여 잠재적 연계(possible links)의 개수가 최소화 되도록 선정한다.

연계(links) : $W >$ 상위 임계점

잠재적 연계(possible links) : 하위 임계점 $\le W \le$ 상위 임계점

비연계(non-links) : $W <$ 하위 임계점

이론적으로는 레코드 쌍 (i, j)가 연계이고 (j, k)도 연계이면, 레코드 쌍 (i, k)는 연계되어야 한다. 그러나, 실제 적용에서는 레코드 쌍 (i, j)와 (j, k)가 모두 연계인데, (i, k)가 비연계인 모순적인 경우가 발생할 수도 있는데, 이런 경우는 거의 발생하지 않을 것으로 추정된다. (i, j)와 (j, k)가 연계라면 이 쌍들의 종합가중치가 확연히 높다는 것을 의미하고, 이는 일치하는 연계변수의 개수가 더 많다는 것을 의미한다. 예를 들어 4개의 연계변수를 사용한다면, (i, j) 또는 (j, k)가 연계라는 것은 각각 3개 이상의 변수가 일치한다는 것을 뜻하며, (i, k)도 최소 2개 변수이상은 일치하게 될 것이다. 그러므로 (i, k)도 연계될 확률이 크다고 할 수 있다. 만약,

Table 3: Contingency table of link status and true match status

True Match status	Link status	
	Links	Non-Links
True Match	True Links	False Non-links
True Non-match	False Links	True Non-links

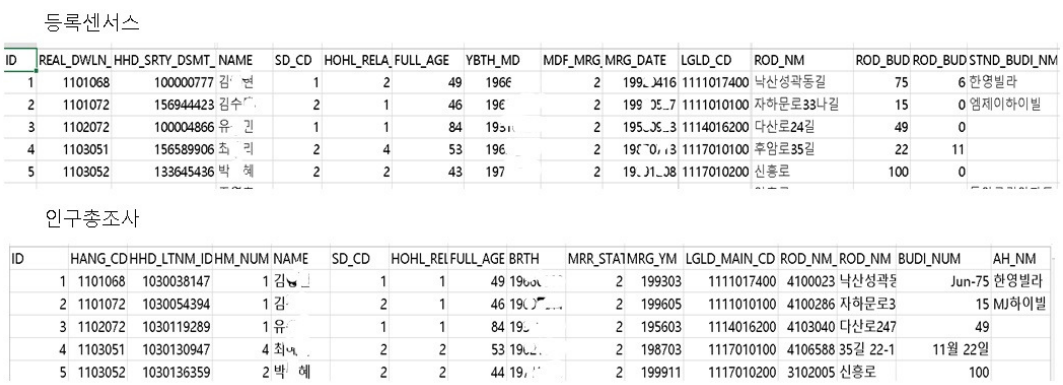

Figure 1: *Data structure.*

연계변수의 개수가 많고 낮은 임계치를 사용할 경우 그런 사례가 발생할 가능성도 있으나, 연계의 에러율이 높아지는 결과를 초래할 것이기 때문에 임계치를 높일수록 모순적인 경우가 발생할 가능성은 줄어들 수 밖에 없다. 본 논문에서 사용한 자료에서는 이러한 모순적인 사례는 발생하지 않았다.

3.5. 연계결과의 품질 평가

연계 결과의 품질은 실제 동일 개체가 연계되었는지에 달려 있다. 만약 실제 매치 상태(true match)를 알 수 있다면 거짓 연계(false links, 동일 개체가 아닌데 연계된 것)와 거짓 비연계(false non-links, 동일 개체인데 연계되지 않은 것)의 개수를 계산할 수 있다. 연계 오류(linkage error)의 계량적 지표로 민감도(sensitivity = True Links/Total True Matches)와 정밀도(precision = True links/ Total links obtained)를 계산함으로써 연계 결과의 품질 평가에 사용할 수 있다. 민감도가 높을수록 정밀도는 낮게 나타나고, 정밀도가 높을수록 민감도는 낮아지는 역의 관계가 존재한다. 따라서 두 지표 간에 균형을 이루는 임계값을 찾는 것이 필요하다.

4. 등록센서스-인구총조사 연계 시뮬레이션

본 절에서는 앞에서 설명한 확률적 자료연계 모형을 적용하여 통계청의 등록센서스와 인구총조사의 표본조사를 결합하는 데 적용해 본다. 통계청의 인구주택총조사는 등록센서스(매년 행정자료를 결합하여 우리나라 인구와 주택의 전수 통계 작성)와 20% 표본조사(5년 주기로 전체의 20% 표본가구를 대상으로 심층항목 조사) 두 가지가 혼합되어 있다. 여기에 사용된 데이터는 등록센서스와 인구총조사 표본조사 자료로 부터 각각 1,005개 레코드를 추출하여 개인정보 보호를 위하여 가공한 것이며, 추출된 양쪽 데이터는 연계결과의 성능을 평가할 수 있도록 같은 개인들의 레코드로 구성하였다. 연계 작업에 사용된 소프트웨어 도구는 R의 `reclin` 패키지이다. 두 데이터는 Figure 1과 같은 형태로 되어 있다.

여러 변수들 중 연계에 활용할 변수는 이름, 성별, 생년월일, 도로명주소, 지역코드이다. 두 데이터 연계를

Table 4: Generation of comparison pairs

	x	y
1	1	1
2	1	2
3	1	3
4	1	4
⋮	⋮	⋮

위해 생성되는 비교 쌍(comparison pairs)은 총 1,010,025(= $1,0051,005$)이며, 이 중 연계될 수 있는 레코드 수는 최대 1,005개이다. 위 자료를 이용하여 두 개의 자료가 연계되는 과정을 각 단계별로 살펴 보도록 한다.

4.1. 표준화(standardization)와 정제(cleaning)

연계 작업을 진행하기 전에 두 파일에 있는 변수들의 필드 내용을 표준화하거나 정제하는 일이 선행되어야 한다. 자료 출처에 따라 주소, 생년월일 등을 표기하는 양식이 각각 다른 경우가 많기 때문이다. 예를 들어, 같은 주소라도 '동이름' 또는 '도로명'으로 다르게 표기될 수 있고, 생년월일도 '0000년00월00일' 또는 '0000/00/00'로 표기방식이 다를 수 있다. 영문의 경우 이름이나 주소에 축약형도 많다. 표준화를 하지 않을 경우 표기방식의 차이로 인해 매칭에 있어서 손실이 커지고 정확도가 하락하게 되므로 자료연계의 성공 확률을 높이기 위해서는 표준화가 반드시 선행되어야 한다. 또한, 결측이 있는 경우 해당 연계변수는 불일치로 판정되어 가중치가 음수가 되기 때문에 연계 가능성이 떨어질 수밖에 없다. 연계의 성공률을 높이기 위해서는 되도록 결측이 없는 변수를 연계키로 사용하거나 사전에 결측을 제거해 주는 것이 필요하다. 본 논문에서 시뮬레이션용으로 사용한 자료는 이미 정제와 표준화를 마친 것이다.

4.2. 비교쌍의 생성

자료연계의 첫 단계는 비교를 위한 레코드쌍을 생성하는 것이다. 두 파일에 속한 동일 개체를 찾아내어 연계시키기 위해서는 두 데이터 세트에 속한 모든 레코드의 조합을 비교하여야 한다. 완벽한 연계키(linkage key)가 있다면 굳이 모든 레코드 조합을 일일이 비교할 필요가 없을 것이지만, 연계키가 부정확하거나 입력오류가 존재하기 때문에 모든 조합을 비교하여야 한다. 이를 위해 두 데이터 세트로부터 레코드의 비교쌍들(comparison pairs)을 생성한다. 두 파일이 각각 1,005개의 레코드를 담고 있으므로 생성되는 비교 쌍의 개수는 $10051005 = 1,010,025$이며, Table 4와 같은 형태로 비교쌍이 생성된다.

4.3. 블로킹(blocking)

현실에서는 연계하고자 하는 데이터의 레코드 수가 몇 십만, 몇 백만인 경우도 많다. 그런 경우에 생성되는 비교쌍의 개수가 어마하게 많아진다. 레코드 수가 일백만이라고 하면 비교 쌍의 개수는 $1,000,000^2$이 된다. 이럴 때 블로킹을 통해 비교 쌍의 개수를 줄여줌으로써 연계 작업의 효율을 높일 수 있다. 블로킹 변수를 설정하면 그 변수의 필드 값이 정확히 일치하는 레코드들만 비교가 이루어진다. 예를 들어, '지역'을 블로킹 변수로 사용하게 되면 같은 지역에 해당하는 레코드들끼리만 비교 쌍이 생성되는 것이다. 블로킹의 주된 목적은 연계에 부적합한 후보 쌍들을 걸러내는 것이다. 성별, 지역, 생년 등을 블로킹 변수로 사용할 수 있고, 몇 개 변수를 혼합하여 사용할 수도 있다. 중요한 것은 최대한 오류가 없는 변수를 사용하는 것이다. 등록센서스-인구총조사 데이터에 '지역'을 블로킹하면 비교 쌍의 개수는 176,821로 줄어든다. 비교쌍의 개수는 블로킹 하지

Table 5: Agreement patterns of comparison pairs

	x	*y*	Name	Gender	Birth Date	Address
1	1	1	TRUE	TRUE	TRUE	TRUE
2	1	2	FALSE	FALSE	FALSE	FALSE
3	1	3	FALSE	TRUE	FALSE	FALSE
4	1	4	FALSE	FALSE	FALSE	FALSE
5	1	5	FALSE	FALSE	FALSE	FALSE

Table 6: Jaro-Winkler similarity score

	x	*y*	Name	Gender	Birth Date	Address
1	1	1	1.0000	1	1.0000	1.0000
2	1	2	0.5556	0	0.6667	0.4306
3	1	3	0.0000	1	0.7778	0.4260
4	1	4	0.0000	0	0.6667	0.4260
5	1	5	0.0000	0	0.7500	0.0000

Table 7: Linkage rate (simple deterministic vs. similarity score)

	Link	Non-link	Linkage rate (link/total)
Deterministic	416	589	41.4
Similarity score	481	524	47.9

않았을 때보다 1/6로 감소한 것이다.

4.4. 연계변수(linking variable) 필드값 비교

‘지역’을 블로킹한 후 비교쌍이 생성되고 나면, 비교쌍을 대상으로 각 연계변수의 필드 값끼리 비교가 시행되어 연계변수 값의 일치/불일치 결과가 산출된다. 연계변수 필드 값이 완벽히 일치하면 TRUE, 그렇지 않으면 FALSE로 나타난다(Table 5). 이것이 결정적 연계방법(deterministic linkage)의 기본 원리이다.

한편, 실제 데이터에서는 오탈자 등으로 인해 동일 개체일지라도 필드 값이 일부만 일치할 수 있다. 따라서 단순히 일치/불일치로만 판단하면 연계에서 누락되는 동일 개체들이 다수 있을 것이다. 이를 보완하기 위해 문자열 비교함수(string comparison function)에서 산출된 유사도 점수(similarity score)를 사용함으로써 동일 개체들의 연계 확률을 높일 수 있다. 문자열 비교함수에 의해 두 필드의 similarity를 0과 1 사이로 점수화하여 similarity가 일정 점수(0.95 또는 0.9) 이상이면 일치, 아니면 불일치로 처리하는 방식이다. 문자열 비교함수로는 Jaro-Winkler, Levenshtein 등이 사용된다. Jaro-Winkler 함수는 Jaro에 의해 처음 소개되고 Winkler에 의해 수정된 것이다 (Winkler, 1990). Jaro는 두 단어에서 일치하지만 순서가 다른 글자들의 위치교환(transposition) 원리를 이용하여 similarity를 계산하는 식을 제시하였으며, Winkler는 여기에다 앞 글자(prefix)가 일치할 때 높은 가중치를 주어 점수를 계산하였다. similarity score는 0과 1 사이의 값을 가지며, 두 단어가 완전히 일치하면 1, 완전히 불일치하면 0의 값을 갖는다. Jaro-Winkler 함수를 적용했을 때 각 필드 점수는 Table 6과 같이 산출된다. 4개의 연계변수가 모두 일치하는 비교 쌍들을 연계해볼 때, 단순 완전 일치/불일치 패턴(Table 5)에 의한 연계율은 41.4%이며, Jaro-Winkler similarity 점수(Table 6)를 활용하여 0.95 이상인 필드가 일치하는 것으로 처리했을 때 연계율은 47.9%로 상승함을 볼 수 있다.(Table7)

Table 8: *m*-probability, *u*-probability and weights

	m-probability	*u*-probability	agreement weight	disagreement weight
Name	0.9932	1.6921e-04	8.6776	–4.9979
Gender	0.9949	4.9664e-01	0.6888	–4.6041
Birth Date	0.7454	7.7121e-05	9.1764	–1.3681
Address	0.5543	1.6378e-02	3.5218	–0.7916

Table 9: Distribution of match weights

(–11.8, –9.35]	(–9.35, –6.93]	(–6.93, –4.51]	(–4.51, –2.1]	(–2.1, 0.32]	(0.32, 2.74]	(2.74, 5.15]
24,842	51,948	11,510	75,156	2,475	9,221	604
(5.15, 7.57]	(7.57, 9.98]	(9.98, 12.4]	(12.4, 14.8]	(14.8, 17.2]	(17.2, 19.6]	(19.6, 22.1]
52	19	94	131	32	187	550

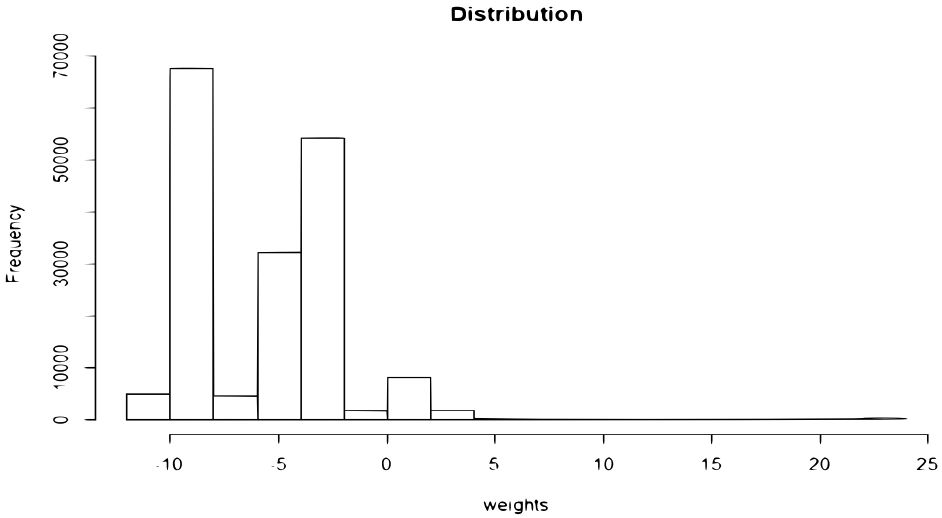

Figure 2: *Distribution of match weights.*

4.5. *m*-확률, *u*-확률, 가중치

각 연계변수의 *m*-확률과 *u*-확률을 얻기 위하여 EM 알고리즘을 이용하였으며 Table 8과 같은 결과가 산출되었다. *m*-확률은 동일 개체일 때 두 레코드의 연계변수 필드 값이 일치할 확률이다. 현실에서는 입력 오류 등이 있으므로 에러율을 반영하고 있다. 이름보다 주소의 *m*-확률이 낮다는 것은 주소의 에러 확률이 더 높다는 것이다. 각 연계변수의 *m*-확률과 *u*-확률이 주어지면 식(3.5)에 의해 일치 가중치와 불일치 가중치를 계산할 수 있으며, 이를 합산하여 모든 비교쌍의 종합가중치가 산출된다. Table9는 종합가중치 구간별 비교쌍 빈도수이며, Figure 2은 비교쌍들의 종합가중치 분포를 그래프로 나타낸 것이다.

4.6. 임계값(threshold) 선정

두 레코드 사이에 연계변수 필드가 서로 일치하면 가중치가 (+)값을 갖고 불일치하면 (-)값을 갖기 때문에 종합가중치가 클수록 동일 개체일 확률이 높아진다. 산출된 종합가중치를 기반으로 연계(links) 또는 비연계(non-links)로 분류하는 임계값(threshold)을 선택한다. 임계값을 어떻게 정하느냐에 따라 연계(link)나 비연계

Table 10: Contingency table of linked pairs

True Match status	Linked pairs	
	Links	Non-Links
True Match	1,003	2
True Non-match	6	175,810

Table 11: Sensitivity and precision by thresholds

Thresholds	9	10	11	12
Sensitivity	99.8	98.4	97.0	91.9
Precision	99.4	99.7	100.0	100.0

Table 12: Performance by selection of linking variables

Linkage variable	Sensitivity	Precision
Name, Gender, Birth Date, Address	99.8	99.4
Name, Gender, Birth Date	99.8	99.2
Name, Birth Date, Address	99.8	98.5
Name, Gender, Address	99.9	99.2
Name, Gender	99.5	99.2
Name, Birth Date	100.0	98.2
Name, Address	95.2	97.5

(non-link)로 분류되는 비교쌍의 개수가 달라지므로 최적의 임계값을 찾기 위한 판단이 필요하다. Table 9에서 종합가중치의 최소값은 –11.8, 최대값은 22.1로 나타나고 있으며, 최상위 두 그룹은 동일 개체일 가능성이 매우 크다. 사용한 데이터에서 연계 가능한 최대 레코드 수가 1,005개임을 감안하여 가중치별 빈도를 합하여 보면 9나 10 정도가 임계값으로 적정해 보인다. 임계값을 9로 선택했을 때 총 1,008개가 연계로 분류되었다. 이 중에 참 연계(true links)는 1,003개이고, 거짓 연계(false links)가 6개이다(Table 10). 본 분석에서 사용한 데이터의 경우 실제 매치 상태를 이미 알고 있기 때문에 연계결과의 성과를 확인해볼 수 있었다. 그리고 임계값을 다르게 선택하여 연계 결과의 민감도(sensitivity)와 정확도(precision)를 측정해 보면 Table 11과 같이 나타났다.

실제 적용에서는 매치 상태를 사전에 모르는 경우가 대부분이기 때문에 연계 결과를 다시 눈으로 확인하고 검증하는 작업이 이루어져야 한다. 예를 들어 한 명에 대해 여러 명이 매치되는 결과가 나올 수도 있다. 이럴 때 종합가중치가 가장 큰 한 명을 연계로 처리하는 방법이 합리적이지만, 경우에 따라 매치되는 여러 명을 선택해 놓고 직접 눈으로 확인하는 추가적인 과정을 거칠 수도 있을 것이다. 확률적 자료연계 방법을 사용하더라도 clerical review 과정은 반드시 필요하다.

4.7. 연계변수 효과

자료연계를 실행할 때 어떤 연계변수를 사용하느냐에 따라 연계의 효용성이 달라질 수 있다. 연계변수의 선택에 따른 연계 결과의 비교를 위해 블로킹 변수(지역)와 문자열 비교함수(Jaro-Winkler)는 고정하고 연계변수 조합을 변경해 가면서 연계를 시도해보았다. 연계변수가 달라지면 m-확률, u-확률, 가중치가 변하기 때문에 연계/비연계 분류를 위한 최적의 임계값도 달라진다. 여기서는 false links와 false non-links의 합이 최소화되

는 임계값을 설정하여 비교하였다. Table 12를 보면, '이름','성별', '생년월일', '주소'라는 변수 중 4개를 모두 연계변수로 사용할 때 가장 효과적인 결과를 보였다. 3개 변수를 사용할 때는 '성별'이 포함되는 것이 더 효과적이었고, 2개 변수만 사용할 때도 '성별'이 연계변수로서의 효용성이 상대적으로 크게 나타났다. 아마도 성별 자료의 경우 오타, 조사 착오 등 오류 확률이 낮기 때문으로 추정된다.

5. 정리의 글

본 논문은 확률적 자료연계가 무엇인지, 어떤 원리로 이루어지는지를 실제 데이터에 적용하여 설명하는 데 주안점을 두었다. 우선 확률적 자료연계 방법의 이론적 토대가 되는 Fellegi-Sunter 모형의 내용과 관련 파라미터(m-확률, u-확률), 연계/비연계 판정 기준 등에 대해 기술하였다. 그리고 등록센서스와 인구총조사 실제 데이터를 활용하여, 확률적 연계방법이 적용되는 과정과 각 단계별로 산출되는 결과들을 살펴보았다. 시뮬레이션 결과를 통해 적절한 연계변수가 확보된다면 확률적 자료연계 방법이 실제 데이터 결합에 충분히 활용 가능하다는 것을 확인할 수 있다. 현재 통계청 등 정부나 공공기관에서 실시하고 있는 데이터 결합 또는 연계는 주민등록번호를 직접 이용하거나 가상키로 변환한 고유키를 매개로 한 결정연계 방식이다. 유일한 키(key)가 있을 때 성공적인 연계 결과를 얻을 수 있으며, 그렇지 않을 경우에는 데이터 결합 시 누락되는 레코드가 상당히 많을 것이다. 간접적 식별정보를 가지고 확률적 연계방법을 활용한다면 고유번호가 없는 경우라 하더라도 데이터 연계의 효율성과 성공확률을 향상시킬 수 있을 것이다.

References

Christen P (2007). A two-step classification approach to unsupervised record linkage. In *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*, **70** 111–119.

Christen P and Goiser K (2007). Quality and complexity measures for data linkage and deduplication, *Quality Measures in Data Mining*, 127–151, Springer.

Dunn HL (1946). Record linkage, *American Journal of Public Health and the Nations Health*, **36**, 1412–1416, American Public Health Association.

Dempster AP, Laird NM, and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**, 1–22, Wiley Online Library.

Elfeky MG, Verykios VS,Elmagarmid AK, Ghanem TM and Huwait AR (2003). Record linkage: A machine learning approach, a toolbox, and a digital government web service,Citeseer.

Fellegi IP and Sunter AB (1969). A theory for record linkage, *Journal of the American Statistical Association*, **64**, 1183–1210, Taylor & Francis.

Feigenbaum JJ (2016). Automated census record linking: A machine learning approach(Working Paper), Harvard University, US.

Goeken R, Huynh L, Lynch TA and Vick R (2011). New methods of census record linking, *Historical methods*, **44**, 7–14, Taylor & Francis.

Hand D and Christen P (2018). A note on using the F-measure for evaluating record linkage algorithms, *Statistics and Computing*, **28**, 539–547, Springer.

Herzog TN, Scheuren FJ, and Winkler WE (2007). *Data Quality and Record Linkage Techniques*, Springer Science & Business Media, New York.

Jaro MA (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, **84**, 414–420, Taylor & Francis Group

Newcombe HB, Kennedy JM, Axford SJ, and James AP (1959). Automatic linkage of vital records, *Science*, **130**, 954–959, JSTOR.

Winkler WE (1990). String Comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceeding of the Section on Survey Research Methods*,US. ERIC.

Winkler WE and Thibaudeau Y (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census(Working Paper), United States Census Bureau.

Winkler WE (1993). Improved decision rules in the fellegi-sunter model of record linkage, **56**, Citeseer

Winkler WE (1995). Matching and record linkage, *Business Survey Methods*, **1**, 355–384, New York.

Winkler WE (2000). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*,US Bureau of the Census Washington, DC.

Received June 24, 2021; Revised July 19, 2021; Accepted July 23, 2021

확률적 자료연계의 이론과 적용에 관한 연구

최연옥[a], 이상인 [1, b]

[a]통계청, 충남대학교 정보통계학과 [b]충남대학교 정보통계학과

요 약

본 논문은 확률적 자료연계 방법의 기본 개념과 이론적 모형을 소개하고, 실제 통계청 데이터를 사용하여 확률적 자료연계가 진행되는 과정과 원리를 보여준다. 먼저 확률적 자료연계와 결정적 자료연계와의 차이를 간단히 알아보고, 확률적 자료연계 방법론의 토대가 되는 Fellegi-Sunter 모형의 기본 구성과 관련된 모수(m-확률, u-확률), 가중치, 매치여부 판정기준에 대해 기술한다. 그리고 통계청 등록센서스와 인구총조사 자료를 이용하여 그 모형을 적용한 자료연계가 이루어지는 구체적인 과정에 대해 설명하고, 이를 통해 얻어진 연계결과의 정확성을 살펴본다.

주요용어: 확률적 자료연계, Fellegi-Sunter 모형, *m*-확률, *u*-확률, 연계가중치

본 논문은 한국연구재단 지원에 의한 논문임(No. 2020R1I1A3071646).

[1]교신저자: 대전광역시 유성구 대학로99, 충남대학교 정보통계학과. E-mail:sanginlee44@gmail.com