*Journal of The Korea Institute of Information Security & Cryptology* ISSN 1598-3986(Print)
VOL.30, NO.3, Jun. 2020 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2020.30.3.503

# 데이터 3법 시대의 익명화된 데이터 활용에 대한 제언

천 지 영,[1†] 노 건 태[2‡]
[1]이화여자대학교 (교수), [2]서울사이버대학교 (교수)

# Suggestions for Applications of Anonymous Data under the Revised Data Privacy Acts

Ji Young Chun,[1†] Geontae Noh[2‡]
[1]Ewha Womans University (Professor), [2]Seoul Cyber University (Professor)

요 약

데이터 3법으로 인해 개인정보를 가명처리 후 데이터를 공개할 수 있게 되었다. 이렇게 익명화된 데이터는 연구 및 서비스 분야 등에서 유용하게 활용될 전망이나, 익명화된 데이터로부터 정보의 주체를 재식별하는 등 프라이버시 침해에 대한 우려가 크다. 본 논문에서는 공공 데이터에서 개인을 식별해내는 것이 크게 어렵지 않음을 보이고, 또한 공개된 데이터의 신뢰성에 의문을 제기한다. 사용자들이 데이터 공개와 프라이버시 보호 사이의 상충관계를 잘 이해하여 데이터 3법 시대에 데이터를 안전하게 활용할 수 있는 방안에 대해 제언한다.

ABSTRACT

The revisions to data privacy acts allows the disclosure of data after anonymizing personal information. Such anonymized data is expected to be useful in research and services, but there are high concerns about privacy breaches such as re-identifying of the individuals from the anonymized data. In this paper, we showed that identifying individuals from public data is not very difficult, and also raises questions about the reliability of the public data. We suggest that users understand the trade-offs between data disclosure and privacy protection so that they can use data securely under the revised data privacy acts.

**Keywords:** Anonymization, Pseudonymization, Public data, Revisions to data privacy acts

## I. 서 론

최근 스마트폰, 웹로그, 사물인터넷, 소셜네트워크 등을 통해서 데이터가 폭발적으로 증가하고 있다. 데이터의 증가량은 기존의 데이터보다 너무 방대하여 현재의 데이터 저장 기술을 넘어서고 있으며, 따라서 이러한 데이터를 저장하고 관리하기 위한 새로운 기술이 요구된다. 또한, 서비스 제공자는 이렇게 생성된 데이터를 분석하고 활용하여 각각의 개인에게 특화된 맞춤형 서비스를 제공할 수도 있게 된다. 하지만 고객 입장에서는 자신의 정보가 상업적으로 이용되는 것이 불편할 수 있으며, 또한 개인의 사생활 노출에 대한 우려가 있을 수 있다. 단적으로 2012년 미국의 대형마트인 타겟(Target)에서 고등학생에게 유아용품 할인 쿠폰을 보내 곤란을 겪었던 사건을 들 수 있겠다[1]. 이 여학생이 영양제를 구입한 후 얼마 되지 않아 로션을 구입했기 때문인데, 빅데이터 분석을 통해 여성이 임신하면 초기에는 영양제를, 중기에는 로션을, 그리고 말기에는 유아용품을 주로 구매한다는 사실을 알았기 때문이다. 타겟에서는 고객의 구매 패턴을 분석하여 맞춤형 서비스를 제공하고자 하였으나, 여학생 입장에서는 임신 사실이 부모에

Received(02. 20. 2020), Modified(05. 04. 2020), Accepted(05. 04. 2020)
† 주저자, jychun@ewha.ac.kr
‡ 교신저자, gnoh@iscu.ac.kr(Corresponding author)

게 감추고 싶었던 비밀이었을 것이다.

빅데이터를 활용하면 기존에 알지 못했던 사용자 개인의 성향을 분석할 수 있게 되고, 이를 활용하여 개인에게 맞춤형 서비스를 제공할 수 있게 되는 반면, 타겟의 사례처럼 이러한 정보가 오용되거나 남용될 경우 사용자의 프라이버시를 침해할 수 있게 된다. 따라서 개인의 프라이버시를 보호하면서도 데이터를 활용할 수 있도록 하기 위한 발판을 마련하기 위해 데이터 3법이 개정되었으며, 이 개정안이 2020년 1월 9일 국회 본회의를 통과하였다. 데이터 공개는 시대적 흐름에서 어쩔 수 없는 선택이며, 안전하게 활용될 수 있는 기술 및 규제가 필요하다는 방증으로 볼 수 있다.

개인정보 보호법 일부개정법률안[2]에서 특히 주목할 부분은 가명정보의 처리에 관한 특례이다. 여기서는 개인정보 처리자가 정보 주체의 동의 없이 통계작성, 과학적 연구, 공익적 기록보존 등을 위해 가명정보를 처리할 수 있다고 되어있다. 미국의 경우에도 의료보험의 이동과 책임에 관한 법률(Health Insurance Portability and Accountability Act; HIPAA)에서 정한 방법에 따라 비식별처리된 환자의 개인정보를 공유하여 의학 연구에 활용될 수 있도록 하고 있다. 하지만 가명처리되거나 비식별처리된 개인정보로부터 다시 정보의 주체를 재식별할 수 있게 된다면 프라이버시 침해문제가 발생할 수 있기 때문에, 안전하게 데이터를 공개할 수 있는 방법이 요구된다.

데이터 3법에 따라 가명처리 후 데이터를 공개하였을 때 개인 프라이버시 침해문제가 발생할 가능성이 매우 높다. 이미 개인정보를 가명처리해서 데이터를 공개해왔던 미국에서도 프라이버시 침해문제가 끊이지 않고 있는데, 다음 장에서 살펴볼 아메리카 온라인 프라이버시 침해 사례라든지 개인 유전체 데이터 공개로 인한 프라이버시 침해 사례 등이 바로 그것이다. 우리나라에서도 공공 데이터를 공개하고 있는데, 이러한 데이터들로부터 사용자의 프라이버시가 침해될 여지가 농후하다.

따라서 본 논문에서는 서울시 공공자전거 이용정보 분석을 통하여 공개된 공공 데이터에서 개인을 식별해내는 것이 크게 어렵지 않음을 보인다. 분석을 위해 서울 열린데이터 광장에 공개된 공공자전거 데이터를 활용하였고, 개인을 식별하기 위해 SNS에 공공자전거 이용에 대한 글을 올린 사람들의 정보를 이용하였다. 그 결과 서울시 공공자전거를 이용하는 사용자의 몸무게, 연령대, 성별 등을 알아낼 수 있었다. SNS에 공공자전거 이용에 대한 글을 올릴 때, 그리고 공공자전거를 이용할 때, 이러한 정보가 노출될 것이라는 사실을 사용자는 인지하지 못했을 것이다. 이와 같이 데이터 공유가 활발해지면 서비스 제공자도, 그리고 사용자도 미처 생각하지 못했던 프라이버시 침해문제가 발생할 수 있게 될 것이다.

데이터를 공유함으로써 얻을 수 있는 장점은 단순히 사용자에게 편리성을 제공하기 위한 것으로부터 개인의 생명과 관련된 것까지 다양하다. 프라이버시 침해문제를 고려해서 안전성을 강화시키면 데이터 공유에 제한을 받게 되고, 데이터 공유에 큰 가치를 둔다면 어느 정도의 프라이버시 침해는 감수할 수밖에 없을 것이다. 실제로 자신의 유전체 데이터를 공개한 사용자들은 프라이버시가 침해될 수 있다는 위험을 잘 알고 있었지만, 개인정보 유출에 대한 걱정보다는 자신의 건강상의 문제를 빨리 치료하는 것을 더 중요시하게 생각했다는 설문 결과가 있었다[3]. 따라서 서비스 제공자는 데이터 공유로 인해 얻게 되는 사용자들의 이익과 이로 인해 발생될 수 있는 프라이버시 침해문제를 잘 분석하여 최대한 개인정보를 보호하면서도 가치있는 서비스를 제공할 수 있도록 노력해야 할 것이다. 사용자 입장에서는 개인의 프라이버시 침해가 발생할 수도 있다는 우려 때문에 무조건적으로 데이터 공개를 꺼리기보다는, 기술의 발전이나 가치있는 서비스를 제공받기 위해 어느 정도는 감수해야 할 수도 있다는 사실을 인지하고, 자신의 데이터 공개를 스스로 결정할 수 있어야 할 것이다.

본 논문의 구성은 다음과 같다: 2장에서는 데이터 3법 개정안에 대해 살펴보고, 프라이버시 침해 사례에 대해 살펴본다. 3장에서는 공공 데이터를 분석하고, 4장에서는 연구 방법 및 절차를 살펴본다. 5장에서는 분석 결과를 토대로 개인을 식별함을 보이고, 6장에서는 데이터 3법 시대에 익명 데이터를 활용하기 위한 제언을 한 후, 7장에서 결론을 맺는다.

## II. 연구의 배경과 문제제기

지금까지 공공 빅데이터를 활용하여 서비스나 예측 모델 등을 만든 실제 사례들은 많은 편이나, 국내에서 공개된 빅데이터를 분석하여 개인을 식별한 연구 결과는 쉽게 찾아보기 어렵다. 실제로 공공 빅데이터를 활용하여 1인당 주거면적을 추정하는 연구[4]나 공공 데이터를 활용하여 스마트시티 고장발견

진단 서비스를 개발하는 연구[5], 개방형 공공 데이터를 활용하여 국내 동절기 일별 최대전력 예측 모델 연구[6, LH19] 등 최근 공공 데이터를 활용한 긍정적인 연구 사례들이 속속 등장하고 있다.

언론의 보도나 연구 결과 등으로 드러나는 공공 데이터 활용의 장밋빛 전망과는 달리, 실제로 공공 데이터를 활용하여 개인의 프라이버시가 노출되는 등의 우려가 있으며, 실제로 이러한 우려에 대한 기술적 방안에 대한 기본 연구는 일정 수준 이상 진행되어 있다.

실제로 데이터베이스에 저장되는 개인의 정보를 보호하고자 하는 방법으로는 크게 암호학적 방법과 비암호학적 방법으로 나눌 수 있는데, 암호학적 방법은 상대적으로 개인의 정보를 보호함에 있어 우수성을 가지지만, 공공 데이터의 공개 측면에서 바라보기에는 실효성 측면에서 부족한 것이 사실이다. 반면, 비암호학적 방법으로는 k-익명성(k-Anonymity)[7], l-다양성(l-Diversity)[8], t-근접성(t-Closeness)[9] 등의 방법부터 시작하여 차별적 프라이버시(Differential Privacy)[10] 등 데이터베이스에 저장되는 개인의 정보들을 비식별화하고자 하는 기술적 방안들에 대한 기본 연구는 꾸준히 진행되며 발전되고 있으나, 이러한 조치들은 개인의 정보를 완벽하게 보호하고 있지는 못한다.

실제 국내에서 제공되는 공공 데이터 분석은 암호학적 방법을 사용하지 않으며, 개인을 직접적으로 드러낼 수 있는 이름, 주민등록번호, 핸드폰번호, 신용카드 등의 정보는 삭제하고, 나이나 주소 등과 같은 정보는 버켓팅하여 제공하고 있는 수준이다. 이것은 기본적인 비식별화 조치를 취한 것이며, 이정도 수준의 비식별화 조치는 해외 분석 결과에 따르면 비식별화된 공공 데이터와 기타 정보들을 결합하여 개인을 식별할 수 있으며, 실제로 식별한 사례들이 연구 결과로 등장한 바 있다[8, 9].

국내에서 공공 데이터는 최근 몇 년간 국가 주도하에 빠른 속도로 증가하고 있다. 다만 아직까지 국내 공공 데이터를 분석하여 개인을 식별한 연구 결과는 찾아보기 힘들며, 데이터 3법 시대를 맞이하여 본 논문에서는 실제 공공 데이터와 온라인 상에서 쉽게 얻을 수 있는 기타 정보를 활용하여 개인을 식별하는 과정을 보이고, 이것의 문제점과 해결 방안을 고민하고자 한다.

### 2.1 데이터 3법 개정안

데이터 3법이란 데이터 이용을 활성화하는 3가지 법률인 “개인정보 보호법”, “정보통신망 이용촉진 및 정보보호 등에 관한 법률(정보통신망법)”, “신용정보의 이용 및 보호에 관한 법률(신용정보법)”을 통칭하는 표현으로, 데이터 이용에 관한 규제 혁신과 개인정보보호 문제를 해결하기 위해 2018년 11월 15일, 데이터 3법 개정안이 발의되었다[11]. 이는 대통령 직속 4차산업혁명위원회의 주관으로 시민단체, 산업계, 법조계, 학계 등 각계 전문가가 참여하여 의견이 반영된 입법조치로, 2020년 1월 9일 데이터 3법 개정안이 국회 본회의를 통과하였다.

데이터 3법 개정안의 주요 내용으로는 데이터 이용을 활성화하기 위해 가명정보라는 개념을 도입하였으며, 개인정보보호 관련 법률의 유사성과 중복성을 정비하고자 법체계를 정비해 “개인정보 보호법”으로 일원화하였고, 데이터 활용에 따른 개인정보를 다루는 처리자의 책임을 강화하였으며, 개인정보 판단 기준을 명확히 하였다.

특히, 데이터 이용 활성화를 위해 EU GDPR(General Data Protection Regulation; 일반개인정보 보호법)을 반영하여 가명정보를 도입하였으며, 이를 통해 개인정보 주체의 동의 없이 통계 작성(상업적 목적을 포함), 과학적 연구(산업적 목적을 포함), 공익적 기록보존 등을 위해 가명정보를 사용할 수 있도록 하였다.

결과적으로 데이터 3법 개정안을 통해 데이터가 전 산업의 가치 창출을 위한 새로운 성장 동력을 확보할 수 있다는 기대효과가 있으며, 새로운 기술이나 제품, 개인 맞춤형 서비스 등의 혁신 서비스 창출 활성화를 기대할 수 있다. 또한, 개인정보 감독기구의 독립성을 확보하여 EU 적정성 평가 승인이 예상되며, 국내 기업이 EU 거주자의 개인정보를 이전할 때 필요한 절차를 면제받을 수 있어서 EU 진출에도 도움을 받을 수 있을 것으로 기대된다.

### 2.2 아메리카 온라인 프라이버시 침해 사례

2006년 8월 4일, 아메리카 온라인(America Online; AOL)은 자사의 검색엔진에서의 검색 기록을 익명화한 후 자신의 웹사이트에 공개하였다[12]. 이것은 3개월 동안 650,000명이 넘는 사용자들의 검색 기록으로 2,000만 건의 검색 키워드를 포

함하고 있었다. AOL은 학교나 연구기관에서 연구를 하는데 도움이 될 수 있도록 검색 기록을 공개하였는데, 사람들의 검색 유형을 분석하여 각각의 고객의 특성에 따라서 추천을 하는 등 개인화된 맞춤 서비스에 대한 기술 개발과 연구를 돕기 위해 좋은 의도를 가지고 공개를 한 것이었다[13]. 검색 기록을 공개할 당시 AOL은 사용자들의 프라이버시 보호를 위해 데이터를 익명화하였고, 각각의 사용자들의 아이디도 의미 없는 숫자로 대체하여 공개하였다[14].

하지만 뉴욕 타임스 기자 Michael Barbaro와 Tom Zeller Jr.는 AOL의 공개된 데이터와 전화번호부 등과 같은 사용 가능한 다른 정보로부터 익명화된 아이디 4417749의 신원을 밝혀냈다[15]. 데이터를 익명화하여 공개하였으나 많은 검색 기록에 사용자를 식별할 수 있는 정보들이 포함되어 있었기 때문이다. AOL은 2006년 8월 7일에 공개한 검색 기록을 자신의 웹사이트에서 지웠으나, 자료는 이미 많이 공유되고 퍼져있었다.

AOL의 사례는 프라이버시를 보호하는 데이터의 익명화가 쉽지 않다는 것을 보여주는 사례이다. 하지만 이러한 데이터가 유용하게 사용될 수도 있으므로 프라이버시 침해문제로 인해 데이터 공개를 꺼리지 않고, 프라이버시를 보호하면서도 데이터 공개가 활발히 일어날 수 있도록 프라이버시를 보호하는 익명화 기술 개발이나 법안이 필요할 것이다[13].

### 2.3 개인 유전체 데이터 공개로 인한 프라이버시 침해 사례

전 세계적으로 유전체에 대한 연구로 얻어진 정보를 활용한 정밀의료가 진행되면서 유전체 데이터 공유에 대한 요구가 커지고 있다. 유전체 데이터를 공개하였을 때 질병에 대한 이해와 치료에 도움이 되고, 개인 건강을 유지하는 관련 사업이 발달하는 등의 장점이 있다[3]. 우리나라도 전국 각지에 사는 한국인 41명의 유전체 정보를 통합한 국민 표준 게놈지도를 공개하였다[16]. 이 데이터는 정밀의료 기술 개발에 활용될 빅데이터로 국민 건강에 크게 기여할 것으로 전망된다.

유전체 데이터 활용은 장점이 있는 반면, 익명화된 공개 유전체 데이터로부터 개인을 식별하는 문제가 발생하기도 한다. 2013년 Science에 게재된 논문에서 유전체 연구에 참여한 사람들의 기증된 유전체 정보와 Ysearch.org, 그리고 USsearch.com과 같은 온라인에 공개된 DNA 정보를 가지고 특정인을 식별해냈다[17, 18]. 이러한 연구결과는 익명화하여 공개한 개인 유전체 데이터로부터 프라이버시가 침해될 수 있다는 것을 보여준 사례이다.

2017년 자신의 유전체 데이터를 확인하고자 하는 사람들을 대상으로 한 연구결과를 살펴보면 대부분의 응답자가 유전체 데이터 공유로 인해 프라이버시가 침해될 수 있다는 위험을 잘 알고 있지만, 개인정보 유출에 대한 걱정보다는 자신의 건강상의 문제를 빨리 치료하는 것을 더 중요시하게 생각했다[3]. 따라서 어떤 상황에서는 데이터 공유로 얻을 수 있는 장점으로 인해 약간의 프라이버시 침해는 감수할 수도 있을 것이고, 데이터 공개와 프라이버시 보호 사이의 균형점을 찾는 것도 필요할 것이다.

## III. 공개 데이터 분석

본 장에서는 서울시 공공자전거 이용정보 분석을 통하여 공개된 공공 데이터에서 개인을 식별해내는 것이 크게 어렵지 않음을 보인다. 이러한 분석을 위해 서울 열린데이터 광장에 공개된 서울자전거 따릉이 데이터를 활용하였다. 이 데이터에는 대여일자, 대여시간, 대여소번호, 대여소명, 성별, 연령대, 이용건수, 운동량, 탄소량, 이동거리, 사용시간 등이 포함되어있다. 이 데이터로부터 개인을 식별할 수는 없으나 데이터들간의 상관관계 분석을 통해 사용자들의 몸무게를 도출해낼 수 있었다. 개인을 식별하기 위해 SNS에 공공자전거 이용에 대한 글을 올린 블로거들의 정보를 활용하였다. 그 결과 해당 블로거의 자료를 공공자전거 데이터에서 식별해낼 수 있었고, 따라서 그 블로거의 몸무게, 연령대, 성별 정보를 알 수 있게 되었다.

또한, 공공자전거 데이터에서 공개된 자료가 정확하지 않은 부분이 있음을 알아내었다. 사용자들의 몸무게를 분석했을 때 5의 배수인 몸무게가 많았고, 그중에서도 몸무게가 65kg인 경우가 비정상적으로 많았다.

### 3.1 서울시 공공자전거 따릉이

건강한 자전거 도시를 만들기 위한 서비스로 서울시에서는 따릉이 서비스를 제공하고 있으며, 자전거를 대여할 수 있는 정류장 형태의 공간인 대여소를 마련하여 서울자전거의 대여와 반납이 무인으로 이루

어질 수 있도록 https://www.bikeseoul.com 사이트를 운영하고 있다. 대여소는 시민들이 이용하기 편리한 장소를 중심으로 설치되어 운영되고 있으며, 서울자전거 이용자는 대여소가 설치된 곳이면 언제 어디서나 장소에 구애받지 않고 자전거를 대여하고 반납할 수 있도록 서비스되고 있다.

### 3.2 데이터 출처 - 서울 열린데이터 광장

서울시에서는 모든 서울시민들에게 서울시와 연계기관이 공개한 공공 데이터를 확인할 수 있도록 서울 열린데이터 광장 사이트 http://data.seoul.go.kr를 운영하고 있다. 해당 사이트에서는 연구, 관리, 서비스 제공 등을 위해서 서울시 시정활동 과정에서 수집된 다양한 데이터에 접근할 수 있으며, 지속적인 업데이트가 진행되고 있다.

서울자전거 따릉이를 이용하고 나면 운행정보(주행거리, 시간), 운동량(소모 열량) 등의 정보를 확인할 수 있는데, 이는 개인이 스스로 확인할 수도 있고, 공공 데이터 서비스를 제공하는 서울 열린데이터 광장에도 주기적으로 업데이트가 되며, 이때 공개되는 정보로는 대여일자, 대여시간, 대여소번호, 대여소명, 대여구분코드, 성별, 연령대코드, 이용건수, 운동량, 탄소량, 이동거리, 사용시간 등으로 구성되어 있다. 이렇게 제공되는 데이터는 연령대코드만이 10살 단위로 구분되도록 일차적인 가공이 되어 있으며, 나머지 정보는 원본 데이터로 고스란히 저장되어 서울 열린데이터 광장에서 누구나 쉽게 확인해볼 수 있다.

## IV. 연구 방법 및 절차

### 4.1 공개된 정보들간의 상관관계 분석

서울 열린데이터 광장에서 제공되는 서울자전거 따릉이 이용자의 정보는 크게 대여 관련 정보(대여일자, 대여시간, 대여소번호, 대여소명)와 개인 관련 정보(성별, 연령대코드, 이용건수, 운동량, 탄소량, 이동거리, 사용시간)로 구별될 수 있으며, 개인 관련 정보 중에서 운동량, 탄소량, 이동거리에 어떠한 상관관계가 있는지에 대해 분석해보았다.

분석 데이터의 출처는 서울 열린데이터 광장이며, 2019년 12월 23일에 업데이트된 "서울특별시 공공자전거 이용정보(시간대별)"에서 가장 최신 자료인 2019년 11월 24일부터 11월 30일까지 일주일간의 데이터를 사용하였으며, 이동거리가 0인 데이터를 제외한 237,616개의 데이터를 가지고 분석하였다. 기본적인 기술통계 분석 결과는 위 〔Table 1〕과 같다.

Table 1. Descriptive Statistics

| | Exe | CO | Dis | T |
|---|---|---|---|---|
| Mean | 142.5 | 1.2 | 5225.8 | 23.4 |
| SE | 0.8 | 0.0 | 29.1 | 0.1 |
| Med | 54.89 | 0.47 | 2040 | 14 |
| Mode | 21.62 | 0.26 | 910 | 5 |
| SD | 392 | 3.3 | 14174 | 27.3 |
| SV | 153695 | 10.8 | 200903001 | 745 |
| Range | 11405 | 67 | 290850 | 805 |
| Min | 0.02 | 0 | 10 | 0 |
| Max | 11405 | 67.5 | 290860 | 805 |
| Count | 237616 | | | |

〔Table 1〕에서 운동량은 Exe, 탄소량은 CO, 이동거리는 Dis, 사용시간은 T, 평균은 Mean, 표준오차는 SE, 중앙값은 Med, 최빈값은 Mode, 표준편차는 SD, 분산은 SV, 범위는 Range, 최소값은 Min, 최대값은 Max, 관측수는 Count로 표기하였다.

동일한 데이터에 대해 상관분석 결과는 아래 〔Table 2〕와 같다. 아래 표에는 놀랍게도 탄소량과 이동거리 사이에 1이라는 상관계수를 가진다는 사실을 알 수 있고, 운동량과 탄소량 사이의 상관계수인 0.982644는 운동량과 이동거리 사이의 상관계수와 정확히 일치한다는 사실을 알 수 있다. 이러한 사실을 기반으로 우리는 운동량, 탄소량, 이동거리 사이에 강한 양의 상관관계가 있음을 명확히 확인할 수 있다.

Table 2. Correlation Analysis

| | Exe | CO | Dis | T |
|---|---|---|---|---|
| Exe | 1 | | | |
| CO | 0.982644 | 1 | | |
| Dis | 0.982644 | 1 | 1 | |
| Time | 0.194992 | 0.199092 | 0.199095 | 1 |

### 4.2 탄소량 계산식 도출

위에서 분석된 서울자전거 따릉이의 탄소량과 이동거리의 상관계수 분석을 통해 얻은 상관계수는 1이었으며, 이는 탄소량과 이동거리가 정비례 관계임을 의미한다. 즉, 탄소량은 서울자전거 따릉이 이용자의 이동거리에 정비례하여 계산되는 일차식이 존재한다는 사실을 확인할 수 있으며, 우리가 찾아낸 식은 아래와 같다.

$$탄소량 = 이동거리 \times 0.000232 \quad (1)$$

## V. 분석 내용 및 결과

### 5.1 운동량 계산식 도출

서울자전거 따릉이를 가입할 때 운동량 계산을 위해 몸무게를 선택적으로 입력하도록 하는데, 이 화면을 통해 운동량을 계산하기 위해서는 이동거리와 함께 몸무게가 사용된다는 것을 알 수 있다. 운동량과 이동거리의 상관계수는 강한 양의 상관관계가 존재하며, 서울자전거 따릉이 이용자의 몸무게 차이로 인해 상관관계가 1보다는 조금 작은 0.982644가 됨을 추정할 수 있다. 따라서 몸무게 값을 다르게 몇 번 기록할 수 있다면, 운동량을 계산하는 식을 어렵지 않게 추정할 수 있음을 알 수 있다.

우리는 운동량을 계산하는 공식을 추정하기 위하여 서로 다른 몸무게를 가진 사용자들의 서울자전거 따릉이 이용정보를 획득하여 운동량이 이동거리와 몸무게에 정비례 관계가 있고, 아래와 같이 계산되는 식이 존재한다는 사실을 확인할 수 있었으며, 우리가 찾아낸 식은 아래와 같다.

$$운동량 = 이동거리 \times 몸무게 \times 0.000396 \quad (2)$$

### 5.2 서울자전거 따릉이 이용자의 몸무게 도출

위에서 도출한 운동량 계산식으로 인해 우리는 운동량과 이동거리를 알면 다음과 같은 관계식으로 인해 서울자전거 따릉이 이용자의 몸무게를 계산해낼 수 있으며, 해당 식은 아래와 같다.

$$몸무게 = 운동량 \div (이동거리 \times 0.000396) \quad (3)$$

위의 식을 사용하면 서울 열린데이터 광장에 공개된 정보인 운동량과 이동거리를 사용해 모든 서울자전거 따릉이 이용자의 몸무게를 얻는 것이 가능하다. 실제로 분석된 데이터의 일부는 아래 〔Table 3〕과 같으며, 이는 2019년 11월 24일 데이터의 초반 30개 데이터이다. 여기서 몸무게는 W, 소수점 둘째 자리에서 반올림한 몸무게는 RW로 표기하였다.

소수점 둘째 자리에서 반올림한 결과를 확인해보면, 대부분의 경우에서 사용자들은 자신의 몸무게를 정수로 기입한 것으로 확인되며, 일부 사용자만이 자신의 몸무게를 소수점까지 정확하게 기입한 것으로 파악된다. 그리고 가입 시 선택 입력 항목으로 제공되는 몸무게를 입력하지 않는 경우, 자동으로 몸무게를 65kg로 판단하여 운동량을 계산하게 되는데, 이로 인해 몸무게가 65kg인 사용자의 수가 다른 몸무게를 가진 사용자에 비해 압도적으로 많음을 확인할 수 있다.

Table 3. Weight Analysis

| No. | Exe | CO | Dis | T | W | RW |
|---|---|---|---|---|---|---|
| 1 | 81.63 | 0.77 | 3330 | 78 | 61.90 | 61.9 |
| 2 | 152.56 | 1.33 | 5750 | 39 | 67.00 | 67 |
| 3 | 65.79 | 0.54 | 2340 | 11 | 71.00 | 71 |
| 4 | 11.45 | 0.14 | 590 | 4 | 49.01 | 49 |
| 5 | 61.47 | 0.46 | 1990 | 8 | 78.00 | 78 |
| 6 | 67.57 | 0.55 | 2370 | 16 | 72.00 | 72 |
| 7 | 30.12 | 0.27 | 1170 | 6 | 65.01 | 65 |
| 8 | 24.35 | 0.19 | 820 | 21 | 74.99 | 75 |
| 9 | 186.31 | 1.42 | 6110 | 31 | 77.00 | 77 |
| 10 | 118.15 | 1.06 | 4590 | 7 | 65.00 | 65 |
| 11 | 72.49 | 0.79 | 3390 | 17 | 54.00 | 54 |
| 12 | 36.83 | 0.29 | 1240 | 8 | 75.00 | 75 |
| 13 | 303.65 | 1.98 | 8520 | 53 | 90.00 | 90 |
| 14 | 93.66 | 0.75 | 3240 | 22 | 73.00 | 73 |
| 15 | 96.87 | 0.93 | 4010 | 24 | 61.00 | 61 |
| 16 | 654.14 | 6.08 | 26220 | 10 | 63.00 | 63 |
| 17 | 41.55 | 0.4 | 1720 | 7 | 61.00 | 61 |
| 18 | 1625.48 | 14.65 | 63150 | 46 | 65.00 | 65 |
| 19 | 144.22 | 1.41 | 6070 | 53 | 60.00 | 60 |
| 20 | 24.2 | 0.22 | 940 | 4 | 65.01 | 65 |
| 21 | 20.04 | 0.21 | 920 | 4 | 55.01 | 55 |
| 22 | 139.04 | 1.06 | 4560 | 27 | 77.00 | 77 |
| 23 | 280.18 | 2.45 | 10560 | 49 | 67.00 | 67 |
| 24 | 67 | 0.65 | 2820 | 30 | 60.00 | 60 |
| 25 | 1178.37 | 10.62 | 45780 | 43 | 65.00 | 65 |
| 26 | 26.25 | 0.24 | 1020 | 6 | 64.99 | 65 |
| 27 | 31.15 | 0.28 | 1210 | 17 | 65.01 | 65 |
| 28 | 50.68 | 0.37 | 1580 | 11 | 81.00 | 81 |
| 29 | 87.52 | 0.79 | 3400 | 22 | 65.00 | 65 |
| 30 | 22.91 | 0.21 | 890 | 3 | 65.00 | 65 |

〔Table 4〕에서는 237,616개의 데이터 중에서 55kg부터 75kg까지 서울자전거 따릉이 이용자의 반올림된 몸무게를 분석하였다. 〔Table 4〕에서 소수점 첫째 자리에서 반올림한 몸무게는 RW(Rounded Weight), 사용자 수는 Users(Number of Users)로 표기하였다. 서울자전거 따릉이 이용자 중 상당수는 55kg, 60kg, 65kg, 70kg, 75kg과 같이 5kg 단위로 자신의 몸무게를 기록하는 경우가 많음을 확인할 수 있으며, 특히 65kg의 몸무게가 많은 것은 가입 시 선택 입력 항목인 자신의 몸무게를 입력하지 않은 사용자의 경우를 포함하기 때문에 다른 몸무게를 가진 사용자들에 비해 65kg의 몸무게를 가진 사용자들이 압도적으로 많은 것을 확인할 수 있다.

위에서 조사된 몸무게 대비 서울자전거 따릉이 이용자 수를 보다 시각적으로 나타내기 위하여 우리는

Table 4. Rounded Weight Analysis

| RW (kg) | Users | RW (kg) | Users |
|---|---|---|---|
| 55 | 6,468 | 66 | 3,210 |
| 56 | 2,803 | 67 | 4,705 |
| 57 | 3,027 | 68 | 7,077 |
| 58 | 4,702 | 69 | 4,167 |
| 59 | 2,267 | 70 | 15,027 |
| 60 | 10,795 | 71 | 3,175 |
| 61 | 2,513 | 72 | 6,526 |
| 62 | 4,376 | 73 | 5,656 |
| 63 | 5,203 | 74 | 4,191 |
| 64 | 3,675 | 75 | 11,165 |
| 65 | 44,866 | | |

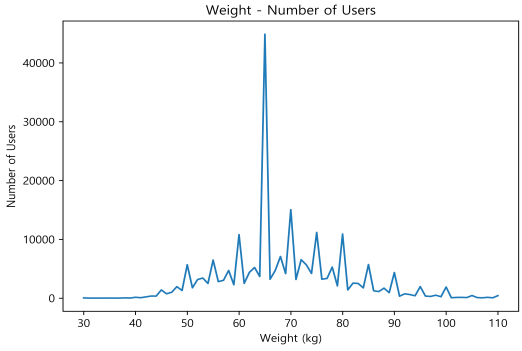


Fig. 1. Graphs for Rounded Weight - Number of Users

Python 3에서 matplotlib 라이브러리를 사용하여 소수점 첫째 자리에서 반올림한 몸무게가 30kg 이상 110kg 이하인 각각의 정수 값들을 〔Fig 1〕과 같이 그래프로 나타내었다. 한눈에 보더라도 사용자들은 자신의 몸무게를 5kg 단위로 어림하여 입력하는 경우가 많음을 확인할 수 있으며, 65kg을 기입한, 또는 선택 입력 항목이기 때문에 입력하지 않은 사용자 수가 다른 몸무게를 가진 사용자들에 비해 압도적으로 많음 또한 확인할 수 있다.

### 5.3 서울자전거 따릉이 이용내역을 공개하는 사람들

서울자전거 따릉이를 이용하는 사람들 중 일부는 자신의 따릉이 이용내역을 자신의 블로그나 SNS 등에 올리는 것을 어렵지 않게 찾을 수 있다.

자신의 서울자전거 따릉이 이용내역을 가공 없이 자신의 블로그나 SNS에 올리는 경우, 위에서 살펴본 바와 같이 원치 않는 자신의 몸무게를 노출하게 되며, 서울 열린데이터 광장에서 해당 이용내역을 찾아서 해당 이용자의 성별이나 연령대 등의 추가 정보를 얻어내는 것 역시 가능하다. 따라서 자신의 이용정보를 올리는 경우, 원치 않는 추가적인 정보 또한 노출될 수 있다는 부분을 사용자 각자가 충분히 인지하고 있어야만 한다.

또한, 자신의 몸무게가 일반적인 사람과는 달리 특이한 경우라면 특히 더 조심할 필요가 있다. 자신의 몸무게와 동일하거나 유사한 몸무게를 가진 서울자전거 따릉이 이용자가 거의 없고, 자신의 활동 반경과 사용 빈도가 어느 정도 노출된 경우라면 어렵지 않게 그 사람의 동선을 파악할 수 있게 된다. 따라서 서울자전거 따릉이 사용자들은 이러한 점을 충분히 인지하고 스스로 조심할 필요가 있다.

### 5.4 서울자전거 따릉이 공개 데이터의 신뢰성

서울자전거 따릉이는 회원가입 시 선택 입력 항목으로 자신의 몸무게를 입력하도록 하며, 입력하지 않는 경우 운동량 등을 계산할 때 65kg으로 계산하게 된다. 우리가 본 논문에서 분석한 결과에 의하면 서울 열린데이터 광장에 올라온 서울자전거 따릉이 사용자들의 몸무게를 도출할 수 있는데, 이 경우 특정 사용자의 몸무게가 정말 65kg인지, 아니면 그 사용자가 선택 항목이기 때문에 입력하지 않아서 나온 결과인 65kg인지에 대한 판단을 하기가 어려워진다.

즉, 자신의 몸무게를 입력하는 것이 회원 가입 시 선택 입력 항목이기 때문에 발생하는 데이터의 신뢰성 문제가 야기될 수 있으며, 신뢰할 수 없는 공공 데이터로 유용한 정보를 생성하기 어려울 수 있다.

또한, 기본적으로 서울자전거 따릉이에서는 사용자의 몸무게를 드러내지 않았으나, 본 논문에서 우리가 도출한 것과 같이 사용자들의 몸무게를 획득할 수 있게 되므로, 서울자전거 따릉이 또는 서울 열린데이터 광장에서 만약 사용자들의 몸무게를 노출하고 싶지 않다면, 지금의 공개 방법과는 다른 방법을 사용할 필요가 있다.

## VI. 시사점과 제언

2장에서 살펴본 아메리카 온라인 프라이버시 침해 사례와 개인 유전체 데이터 공개로 인한 프라이버시 침해 사례, 그리고 3장부터 5장까지 우리가 분석한 서울자전거 따릉이 사용자의 몸무게 도출 및 성별과 연령대 등의 추가 정보를 얻을 수 있는 실제 사례 분석으로부터 데이터 3법 시대에 개인정보의 비식별 처리만으로는 개인의 프라이버시를 보호하는 데에 충분치 않음을 확인하였다. 이는 이미 개인정보를 비식별 처리해서 데이터를 공개해왔던 미국에서도 프라이버시 침해문제가 끊이지 않음으로도 확인할 수 있는 사례이다.

우선 본 논문에서 우리는 공공 데이터로 활용될 수 있는 서비스들에 대해 개인정보 주체에게 이를 확실히 인지시킬 수 있도록 해야 한다고 믿으며, 특정 개인들이 기술의 발전이나 가치 있는 서비스를 제공받기 위해서라면 어느 정도 자신의 프라이버시 침해 문제를 감수할 수 있음을 확인한 바 있다[3]. 따라서 단순히 개인정보의 비식별화, 가명처리 등으로 일관성있게 처리하는 방법이 아니라, 개인의 유전체 공개와 같이 사람의 건강이나 생명과도 직결될 수 있는 부분이라면 어느 정도의 프라이버시 침해를 감수할 수 있다든지, 그리고 단순히 조금의 불편함만을 감소시키는 서비스를 위해서라면 데이터 공유를 어느 정도 제한한다든지 하는 문제를 본격적으로 고민해볼 필요가 있다.

정리하면, 목적에 따라 개인은 자신의 프라이버시 침해문제에 대해 관대하기도, 엄격하기도 하며, 주로 개인의 건강이나 목숨과 관련된 일에 대해서는 개인정보의 침해가 발생하더라도 기술의 발전을 우선시하는 경향이 있다. 따라서 이러한 경우에는 다소 개인의 프라이버시 침해문제가 발생하더라도 이를 보완해줄 법률적/정책적 제도와 규정의 보완이 필요하며, 이 경우 매우 적극적으로 데이터 활용에 방해가 되는 규제를 완화시킬 필요가 있다.

반면, 단순히 사용자들이 약간의 불편함을 감소하기 위해 사용하는 서비스로부터 개인정보가 침해될 수 있는 부분은 사용자들이 받아들이기 어려운 부분일 수 있으므로, 활용의 목적에 따라 개인정보의 비식별화, 익명화, 가명정보 처리 등의 수준을 달리할 필요가 있겠다. 이를 위해 시민단체, 산업계, 법조계, 학계 등 각계 전문가가 참여하여 데이터 활용 목적의 수준을 결정하고, 그에 따른 차별적 개인정보의 비식별화 처리 및 가명정보의 처리 등을 제언한다.

데이터 3법은 이제 개정되어 국회 본회의를 통과하였다. 하지만 우리가 얻고 참고말한한 실제 사례는 미국 등의 외국 사례, 그리고 국내 공공 데이터 제공 서비스의 사례로부터도 충분히 얻을 수 있다. 적합한 목적 달성을 위해 데이터의 적극적인 활용이 필요하나, 그렇지 않은 경우에는 개인정보의 프라이버시 보호 측면이 우선시되는 모습도 바람직하다.

마지막으로, 공공 데이터의 신뢰성 부분에서 검증과 보완이 필요하다. 공공 데이터는 단순히 데이터를 수집하여 제공하는 방대한 데이터이기만 해서는 곤란하다. 공공 데이터로부터 데이터를 분석하고자 하는 사용자들은 그 데이터를 믿고 다양한 형태로 분석하여 제품 및 서비스를 제공하고자 하는데, 앞서 4장에서 분석한대로 공공 데이터의 신뢰성에 문제가 있다면, 신뢰 수준이 낮은 공공 데이터로는 유용한 정보를 생성하기에 한계가 뚜렷할 수밖에 없다. 따라서 공공 데이터를 수집하는 단계서부터 데이터의 품질을 향상시킬 방안을 고민하고, 이를 반영하여 데이터를 수집할 필요가 있다. 조금 더 구체적으로는, 서울자전거 따릉이의 경우 65kg인 사용자가 선택 입력 항목을 입력하지 않은 사용자인지, 아니면 정말 65kg인 사용자인지를 구별할 수 있다면, 보다 정확한 활용이 기대될 수 있으므로, 데이터 수집 단계에서 신뢰성을 높일 수 있는 방안에 대한 고민이 필요하다.

## VII. 결 론

데이터 이용의 활성화를 위한 데이터 3법 개정안의 국회 본회의 통과로 인해 개인정보를 가명처리 후 데이터를 공개하여 통계 작성, 과학적 연구, 공익적 기록보존 등으로 유용하게 활용될 전망이나, 익명화

된 데이터로부터 정보의 주체를 재식별하는 등 개인의 프라이버시 침해에 대한 우려가 높다. 본 논문에서는 공공 데이터로부터 개인을 식별할 수 있음을 서울 열린데이터 광장에 공개된 서울자전거 따릉이 정보로부터 가능함을 보였고, 공개된 데이터의 신뢰성에 의문을 제기하였다. 마지막으로, 데이터 공개와 프라이버시 보호 사이의 상충관계를 이해하여 본 논문에서 제언한 내용이 데이터 3법 시대에 현명한 데이터 활용을 위한 발전 방향에 도움이 되기를 기대한다.

## References


[1] "The credit card company knows my mind that I don't know", http://www.viva100.com/main/view.php?key=20150603010000782, 2015.06.03.

[2] "Personal Information Protection Act", http://www.law.go.kr/lsInfoP.do?lsiSeq=195062&efYd=20171019#0000

[3] The Pros and Cons of Revealing Personal Genomics Data, Sung Hye Hong, BRIC View, 2018.

[4] J-B. Lim and S-H. Lee, "A Study on Estimating Housing Area per capita using Public Big Data - Focusing on Detached houses and Flats in Seoul -," Journal of the Korean Regional Science Association, vol. 36, no. 1, pp. 51-67, Mar. 2020.

[5] H-M. Cho, C-Y. Park, J-H. Kim, and H-I. Jang, "Utilizing Public Data of Building Energy for Fault Detection and Diagnostic Services in Smart City," Journal of Korean Institute of Architectural Sustainable Environment and Building Systems, vol. 13, no. 6, pp. 599-608, Dec. 2019.

[6] G-C. Lee and J. Han, "Forecasting the Daily Peak Load of South Korea During the Winter Season : A Case Study on Open Public Data Usage," J5o8u(1rn0a pl aogf eths)e Korean Operations Research and Management Science Soc, vol. 44, no. 4, pp. 49-58, Nov. 2019.

[7] P. Samarati and L. Sweeney, "Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression," Technical Report, 1998.

[8] A. Machanavajjhala, J. Gehrke, and D. Kifer, "l-Diversity: Privacy Beyond k-Anonymity," ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, Mar. 2007.

[9] N. Li, and T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," Proc. IEEE 23rd International Conference on Data Engineering, May 2007.

[10] C. Dwork, "Differential Privacy," International Colloquium on Automata, Languages, and Programming, LNCS 4052, pp. 1-12, Jul. 2006.

[11] "Revised Data Privacy Acts", http://www.korea.kr/special/policyCurationView.do?newsId=148867915

[12] "AOL search data leak", https://en.wikipedia.org/wiki/AOL_search_data_leak

[13] "Protecting user privacy", http://www.dt.co.kr/contents.html?article_no=2010071502012251697027

[14] Uncharted: Big Data as a Lens on Human Culture, Erez Aiden and Jean-Baptiste Michel, 2013.

[15] The New York Times, "A Face Is Exposed for AOL Searcher No. 4417749", https://www.nytimes.com/2006/08/09/technology/09aol.html, 2006.08.09.

[16] "The Korean Reference Genome Project", http://koreanreference.org/home/index.html

[17] M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, and Y. Erlich,

"Identifying Personal Genomes by Surname Inference," Science, vol. 399, no. 6117, pp. 321-324, Jan. 2013.
[18] The Scientist, "Anonymity Under Threat", https://www.the-scientist.com/daily-news/anonymity-under-threat-39917, 2013.01.17.

## 〈저자소개〉

천 지 영 (Ji Young Chun) 종신회원
1997년 2월: 이화여자대학교 수학과 졸업
2006년 2월: 고려대학교 정보보호학과 석사
2011년 8월: 고려대학교 정보경영공학과 박사
2011년 9월~2019년 12월: 고려대학교 정보보호연구원 연구교수, 시간강사
2012년 8월~2014년 3월: University of Illinois at Urbana-Champaign 박사후 연구원
2020년 1월~현재: 이화여자대학교 엘텍공과대학 컴퓨터공학전공 특임교수
〈관심분야〉 정보보호 프로토콜, 데이터 보안, 프라이버시 향상 기술

노 건 태 (Geontae Noh) 종신회원
2008년 2월: 고려대학교 산업시스템정보공학과 졸업
2010년 2월: 고려대학교 정보경영공학과 석사
2014년 8월: 고려대학교 정보보호학과 박사
2014년 9월~2017년 2월: 고려대학교 정보보호연구원 박사후 연구원, 연구교수
2017년 2월~현재: 서울사이버대학교 빅데이터·정보보호학과 조교수
2020년 3월~현재: 서울사이버대학교 빅데이터·AI센터 센터장
〈관심분야〉 암호 이론, 데이터 보안, 프라이버시 향상 기술