東洋學 第79輯(2020年 4月) 檀國大學校 東洋學硏究院

데이터역사과학

-고대 문명의 보편 질문을 향한 관문-*

김 광 림**

국문초록

오늘날 과학계에서는 역사학과는 다른 방법으로 역사 주제를 연구하고 있으며, 의미 있는 성과를 발표하고 있다. 이 논문은 그러한 연구의 배경에 신뢰성 높은 역사 데이터 구축과 엄밀한 데이터 분석이 공통으로 자리 잡고 있음을 보이며, 데이터과학의 핵심 요소가 포함된다는 점을 설명한다. 이를 위해 유명 자연과학 학술지에 실린 역사학 및 고고학 주제의 논문들을 분석하여, 데이터 구축부터 분석에 이르기까지 데이터과학이 역사 연구에 어떻게 활용되었는지를 검토한다. 이어 과거의 특정 상황을 다양하게 변경하여 시간에 따른 변화를 관찰할 수 있는 '데이터 기반 시뮬레이션' 방법론을 설명하고, 고대 문명 연구에 특히 유용할 수 있음을 강조한다. 학제간 협력이 필수인 대규모의 글로벌 연구 프로젝트에서도 데이터과학이 연구 방법뿐만 아니라 소통의 도구로도 필요함을 보이고, 이를 위한 방법론으로 역사학과 데이터과학의 협력을 전제로 한 '데이터역사과학(Data History Science)'라는 분야를 새로이 제안한다.

[주제어] 역사학, 데이터과학, 데이터역사과학, 협력연구, 데이터베이스, 데이터분석, 고고학, 복잡계 시스템, 시뮬레이션, 고대사, 학제간 연구, 디지털인문학

목 차

* 이 논문은 2019년 제49회 동양학 국제학술대회의 발표문을 수정, 보완한 것이다.

** 단국대학교 사학과 석박사통합과정 / ghim@dankook.ac.kr

Ⅰ. 머리말

고대 연구의 현안을 다룬 한 편의 논문이 2015년『네이처』에 실렸다. 전 세계 25개 기관 소속 39명의 분자생물학, 유전학, 의학, 언어학, 인류학, 고고학 전문가들이 제출한 이 논문은「초원지대에서 시작된 대규모 이주가 유럽의 인도유럽어의 근원이었다.」[1]이라는 제목으로 현재까지 501회 인용되었다.[2] 2019년『사이언스』에 게재된「남부, 중부 아시아의 인구집단 형성」[3]이란 논문도 위의 연구와 비슷하게 고대인의 DNA를 분석하고 추적한 것으로 81개 기관에 속한 117명의 연구자가 공동 저자로 참여했다.

두 논문 모두 고고학의 연구 주제를 탐구하고 있지만, 사실 이 주제는 학문의 범위를 넘어서 오래전부터 인류가 궁금해하던 보편적 질문이다. 이에 과학이 해결사로 나서기 시작하면서 '인류 사회의 기원과 이동'이라는 퍼즐은 더욱 정교하게 맞춰지고 있다. 오랜 역사를 지닌 이 난제를 해결하고자 많은 분야의 연구자들이 인류학적 탐구와 고고학적 증거 수집으로 합리적인 가설을 도출해왔다. 최근에는 분자고고학·분자유전학 형제가 그 가설을 검증하여 이론으로 한 걸음 더 나아가게 만들고 있다.

또한, 이 연구들은 인류의 근원적인 난제를 초학제간 협력과 국제적 협업으로 어떻게 극복할 수 있는지 보여주는 유의미한 사례지만, 역사연구자에게 아쉬움이 남는 결과이기도 하다. 스스로 차려놓은 밥상에 오히려 초대를 받아 함께한 셈이니 말이다. 두 연구 모두 데이비드 앤서니(David Anthony)의 고고학 연구 성과인 '초원 가설'[4]을 중요한 뼈대로 삼고 있으며 연구 결과 또한 이를 지지하는 강력한 증거를 시사한다. 그러나 앤서니를 비롯한 고고학자들의 이름은 공동 연구자 목록 뒤편에 위치한다. 주연은 어디까지나 분자생물학자들이다. 인문사회학이 멋진 음식을 만들어 놓고 연회의 주관은 과학에 맡겨도 괜찮은가? 과학의 향연에 열심히 보조를 맞춰가며, 거대한 프로젝트에 참여하고 있다는 사실만으로도 만족할 수 있는가? 이러한 주도권 집착은 시대착오적인 행태로 비판받기 충분하다. 하지만 역사학이 역사 연구의 충실한 조연으로 자리매김 중이라는 것 또한 엄연한 현실이다.

이 글은 이러한 현실을 인정하고, '합리적 이기주의자'의 자세로 이러한 흐름을 상세히 살펴보려는 취지로 기획되었다. 이를 위해 '최신 연구 동향'과 '자연과학'이라는 키워드로 대표되는『네이처』와『사이언스』를 비

---

1) Haak, Wolfgang · Lazaridis, Iosif · Patterson, Nick · Rohland, Nadin · Mallick, Swapan · Llamas, Bastien · Brandt, Guido, et al., "Massive migration from the steppe was a source for Indo-European languages in Europe.", *Nature* 522, 2015, pp. 207~211.

2) Nature 집계 기준 (2020년 2월 19일 기준)

3) Narasimhan, Vagheesh M. · Patterson, Nick · Moorjani, Priya · Rohland, Nadin · Bernardos, Rebecca · Mallick, Swapan · Lazaridis, Iosif, et al., "The formation of human populations in South and Central Asia.", *Science* 365: 6457, 2019.

4) 초원 가설이란 유라시아 초원지대, 그중에서도 흑해와 카스피해 사이의 북쪽 초원에 살던 사람들이 유럽에 기마와 언어를 전파했으며 이들이 오늘날 인도유럽어족의 조상이라는 가설을 말한다. '쿠르간 가설(Gimbutas, 1956)'로 시작되어 최근에는 '초원 가설(Anthony, 2007)'로 구체화 되었다. 자세한 내용은 다음을 참고:
데이비드 W. 앤서니(저), 공원국(역),『말, 바퀴, 언어: 유라시아 초원의 청동기 기마인은 어떻게 근대 세계를 형성했나』, 에코리브르, 2015(원저 : Anthony, David W., *The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world.*, Princeton University Press, 2010.).

롯한 유명 자연과학 학술지를 중심으로 유사한 연구사례를 모아보고자 한다. 자연과학과 역사학, 또는 고고학이 결합된 최근의 논문을 분석하여 어떠한 공통 배경과 속성을 지니는지를 각 연구 사례별로 살펴볼 것이다. 가급적 고대사와 관련성이 높은 사례를 선정하여 고대 문명 연구에 어떻게 활용될 수 있는지 또한 검토할 것이다. 이러한 연구의 기저에는 역사학 자료 및 사료의 디지털화와 적극적인 데이터베이스 활용이 자리잡고 있다. 이를 종합한 데이터과학(Data Science)의 방법론이 많은 연구에서 어떻게 적용되어 있는지도 들춰볼 것이다. 나아가 공학의 무기 중 하나인 시뮬레이션 방법론을 소개하고, 이를 고대 문명 연구에서 효과적으로 사용할 수 있음을 설명하고자 한다.

마지막으로 서두에 소개한 '대규모 · 국제적 · 초학제간 · 협력' 연구의 배경을 살펴보고 그 미래를 전망하며, 데이터과학과 역사학의 협력이 이러한 흐름에 어떻게 적극적으로 참여할 수 있을지 정리할 것이다. 특히 고대사 연구에 많은 도움을 줄 수 있는 분야로서 데이터과학의 이점을 다시 확인하고 새로운 방법론적 대안으로 '데이터역사과학'을 조심스럽게 제안하고자 한다.

Ⅱ. 데이터 구축: 모든 것의 시작점, 데이터베이스

데이터과학에서 널리 쓰이는 'Garbage-In Garbage-Out(GIGO)'이라는 핵심 명제는 연구 방법론이 훌륭해도 잘못된 데이터가 사용되면 그 결과 전체를 신뢰할 수 없게 된다는 의미이다.[5] 이처럼 데이터 구축은 해당 분야의 전문성을 갖고 과거의 연구를 종합할 수 있어야 가능한, 매우 높은 수준의 작업이다. 그러나 데이터 구축은 개별 연구의 초기 단계에서 해당 연구에 한정된 것만 생산하는 경우가 일반적이다. 데이터 자체를 만드는 연구는 보통 데이터베이스 구축사업으로 진행하기도 하나, 그 자체를 학문적 성과로 높이 평가하지는 않는 경우가 많다. 이렇게 구축된 데이터베이스가 활발히 사용되는 경우가 인문사회학의 경우에는 극히 제한적인데 이러한 흐름은 노력에 비해 결실이 적은 이 과정을 더욱 기피하게끔 만들고 있다.

해외에서는 데이터의 품질과 사용성을 높여 다양한 곳에서 활용할 수 있는 실용적인 방안이 등장하고 있다. 훌륭한 데이터 구축과 그 과정에 대한 연구만을 주제로 다루는 학술지도 활발히 등장하고 있는데,[6] 『네이처』에서도 이러한 흐름에 동참하여 새로운 자매지인 『사이언티픽 데이터(Scientific Data)』를 신설하였다.[7] 이는 2014년에 시작된 오픈 액세스 저널로 매년 빠른 성장을 보이고 있으며 다른 자매지 못지않게 높

5) Corrales, David Camilo · Ledezma, Agapito · Corrales, Juan Carlos, "A conceptual framework for data quality in knowledge discovery tasks (FDQ-KDT): A Proposal.", *Journal of Computers* 10: 6, 2015, pp. 396~405.

6) Candela, Leonardo · Castelli, Donatella · Manghi, Paolo · Tani, Alice, "Data journals: A survey.", *Journal of the Association for Information Science and Technology* 66: 9, 2015, pp. 1747~1762.

7) 『사이언티픽 데이터』는 간결하면서도 핵심만을 담은 설명을 '원칙 Principles'에서 보이고 있다(이하 인용은 Scientific Data 홈페이지, https://www.natureasia.com/ko-kr/scientificdata/ 검색일: 2020.02.17.).
"사이언티픽 데이터는 과학적으로 가치 있는 데이터셋에 대한 기술 및 과학 데이터의 공유와 재사용을 진전시키는 연구를 출간하는 온라인 전용 오픈 액세스 저널입니다. 해당 저널의 주요 콘텐츠 유형인 데이터 디스크립터는 기존의 서술 내용을 구조화 및 큐레이션 기술(메타데이터)과 결합하여 데이터 공유 및 재사용에 새로운 뼈대를 제공합니다."

은 영향력을 지니고 있다.

연구 데이터베이스는 각 대학이나 연구기관에서 개별적으로 공유하는 경우가 일반적이다. 그러나 최근 유럽과 아시아에서는 개별 대학 연구실에서 다루기 어려운 규모의 연구 데이터를 국가 차원에서 구축하기 시작했고, 각 연구기관들이 연합한 비영리단체가 거대 데이터베이스를 제공하는 민간 차원의 활동 또한 미국과 유럽을 중심으로 나타나고 있다.[8] 활발한 연구가 이어지는 데이터베이스들은 모두 월등한 품질과 함께 사용하기 쉽다는 특징을 공통으로 가지고 있다. 이러한 데이터를 이용한 논문이 권위 있는 학술지에 활발히 게재되면서 각 데이터베이스의 위상도 덩달아 오르고 있는데, 이는 데이터 구축과 연구 사이의 선순환 구조를 보다 견고하고 지속적이게 만드는 동인이다. 이러한 경향을 다음 세 논문(A1 · A2 · A3)의 사례 분석을 통해 구체적으로 살펴보도록 한다.

(A1) 「기원전 3700년부터 기원후 2000년까지 세계 도시화의 공간적 특성」[9]

이 연구는 기존의 저명한 도시화 연구 결과를 종합한 것으로, 약 6,000년에 이르는 세계의 도시화 정도를 정량화하여 당대 도시의 위치와 시간에 따른 인구 변화를 보여준다. 주된 자료는 역사학자 터쉬어스 챈들러(Tertius Chandler)[10]와 정치학자 조지 모델스키(George Modelski)[11]의 방대한 저작으로 현대 이전의 도시 이름과 특정 연도의 인구 추정치, 그리고 출처까지 망라하여 수록되어 있다. 이 논문은 비디지털 자료를 디지털 데이터베이스로 구축할 때 필요한 대부분의 작업을 단계별로 잘 정리하고 있는데, 과정 전체는 〈그림 1〉에서 확인할 수 있다. 몇 가지 주목할 만한 특징을 보자면, 종이로 출간된 옛 서적을 전자화한 과정을 상세히 기술한 것이 그 첫 번째로, 다양한 광학인식 기술로 자동화한 부분에 한계점이 있음을 보이며 결국 사람이 어떻게 작업하였는지 설명한 부분이 인상적이다. 두 번째는 서로 다른 두 개의 자료를 하나로 합친 과정에 대한 것으로, 통합 과정에서 의미를 상실할 위험성을 어떻게 최소화하였는지 설명하고 있다.[12] 셋째는 고대 도시의 위치를 지리정보시스템(GIS) 기술을 이용하여 디지털화한 것이다. 이 과정에서

여기서는 과학 데이터로 설명하고 있지만 허용되는 데이터는 '자연과학' 분야에 한정되지는 않는다. "사회 과학, 특히 자연 과학과 사회 과학의 전통적인 경계를 넘어서는 통합 분석에 사용될 수 있는 양적 데이터셋에 대한 설명도 기꺼이 고려할 것입니다."라는 설명이 있고, 이 글에서 다룰 논문도 역사학과 정치학 연구 데이터를 다룬 사례이다.

8) 이혜림, 「국가 고고학 데이터 디지털 아카이브 개발을 위한 연구」, 『한국기록관리학회지』 18: 2, 2018.

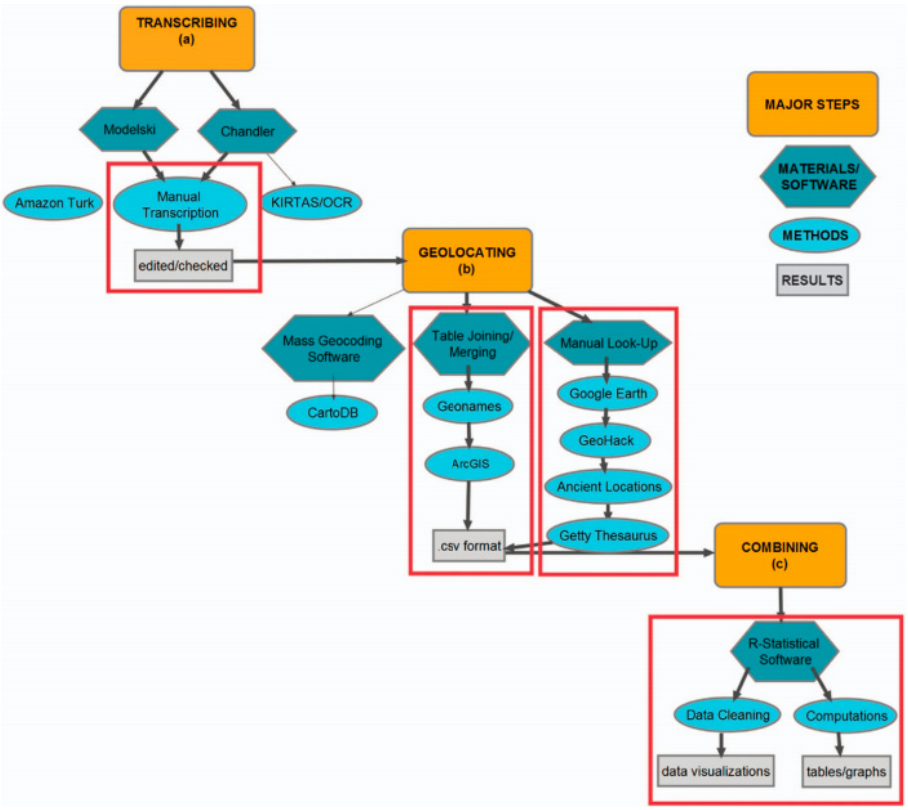
9) Reba, Meredith · Reitsma, Femke · Seto, Karen C., "Spatializing 6,000 years of global urbanization from 3700 BC to AD 2000.", *Scientific data* 3, 2016.

10) Chandler, Tertius, *Four thousand years of urban growth: an historical census*, The Edwin Mellen Press, 1987.

11) Modelski, *George, World cities: -3000 to 2000*, Faros, 2000.

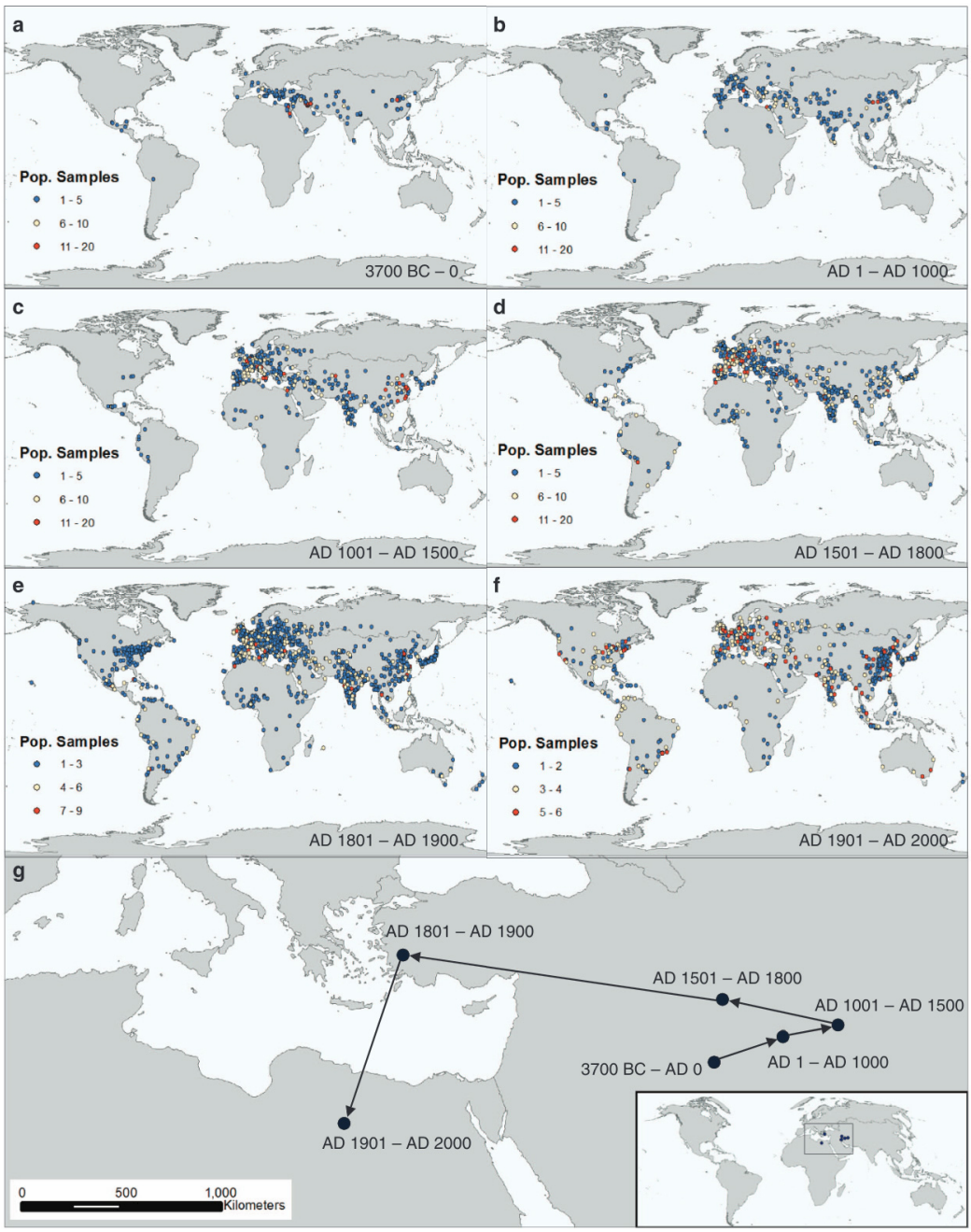
12) 도시공학자인 논문 저자들이 역사적 관점과 정치공학적 관점으로 기술된 연구 자료를 다루는 것은 상당한 한계가 있을 수 있다. 그런 점에서 다른 연구에서 구축된 데이터를 이용해도 괜찮을지 의구심이 들 수 있으나 논문의 저자들 역시 이를 인정하며 최대한 그 문제에 대해 해명하고 있다. 우선 원자료 자체가 많은 후속 연구들로 이어진 것으로 각 분야에서 이미 잘 알려졌다는 사실을 강조하고 있다. 그리고 두 자료를 합치면서 애매 하거나 불일치가 보이는 부분에 대해서는 합리적인 기준을 세우고 확실한 부분만 처리하여 잘못될 여지를 최소화하고 있다. 이 과정 자체를 상세한 설명과 함께 소스코드로 그 논리를 명확히 보여 후속 연구에서 쉽게 정정할 수 있는 여지를 남긴 것 또한 투명하고 확장 가능한 연구의 장점이라 할 수 있다. 이처럼 서로 다른 목적으로 수집된 데이터를 유기적으로 통합하는 방법론은 데이터과학에서 중요하게 다루는 주제로, 그 과정에서 엄격한 기준을 세우고 통계 기법을 이용해 다양한 방식으로 검증하는 것이 핵심이다.

고대 도시의 위치 추정 연구 결과를 모은 데이터베이스와 다양한 지도제작 애플리케이션을 활용한 것 또한 눈에 띄는 점이다.



〈그림 1〉 (a) 서로 다른 원 사료(Modelski, Chandler)의 디지털화 과정. 비디지털 자료를 자동화된 기술(광학인식기술 : OCR)로 만드는 데 한계가 있어, 최종적으로는 수작업을 통해 디지털화와 검수(붉은색 사각형 강조)를 마쳤다.
(b) 과거 도시들의 위도와 경도를 찾아내는 과정. 다양한 지리정보시스템(GIS) 소프트웨어를 이용하였다.
(c) 통계소프트웨어(R)를 이용해 데이터 정제 및 통합 과정과 시각화 작업을 수행하였다.
(출처 : 논문 A1의 Figure 3.)

이렇게 과거 역사에 존재했던 전 세계의 도시에 대해 시간 변화와 공간에 따른 인구 데이터를 구축한 뒤, 그 데이터가 어떻게 의미가 있는지를 다양한 데이터 분석과 시각화를 통해 입증하고 있다. 일례로 〈그림 2〉는 세계 지도에 시간대에 따른 도시 분포와 인구 중심지 이동을 표현한 효과적인 시각화의 결과이다. 연구자들은 이러한 모든 과정을 논문에서 상세히 설명하고 있으며, 연구 전체를 재연하고 검증할 수 있도록 온라인에 소스코드를 공개하였다. 마찬가지로 데이터베이스 자체도 누구든지 활용할 수 있도록 저널과 연구기관의 홈페이지에 게재하였다. 이 논문이 실린 『사이언티픽 데이터』도 이러한 데이터 구축 연구의 흐름을 파악하기 위해 향후 지속적으로 주목할 필요가 있다.



〈그림 2〉 구축된 도시 데이터의 시각화. (a-f) 총 6개의 시대로 구분하고, 각 시대별 도시 인구 데이터가 있는 경우는 점으로 표시했다. 점의 색상과 숫자는 도시별 인구 데이터가 나타나는 빈도를 의미한다.
(g) 한 지역에서 인구 중심지가 시대에 따라 이동했음을 나타내는 지도. 점은 시대별 전체 도시와 인구의 평균값에 해당하는 위치를 의미한다.(출처 : 논문 A1의 Figure 5.)

(A2) 「인간 사회 조직의 전역적 변화를 구조화하는 단일 차원의 복합성을 밝히는 정량적 역사 분석」[13]

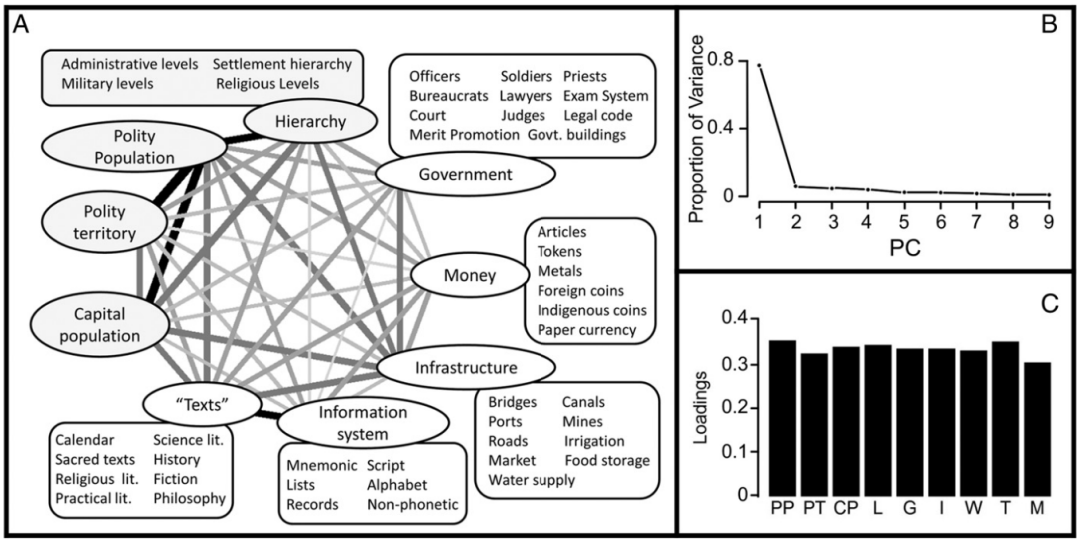
이 논문은 '세스헤트(Seshat)[14]'라는 역사 데이터베이스를 사용하여 전 세계의 여러 문명이 사회적 복합성을 어떻게 증대시켜 왔는지 확인한 연구로, 이렇게 구축된 데이터를 이용해 의미 있는 역사 해석이 가능한지를 입증하려는 시도이다. 세스헤트는 2011년에 유럽과 미국의 인문사회학자들이 결성한 연구 프로젝트의 이름이자 그 결실인 데이터베이스 이름이기도 하다. 복잡한 사회의 진화를 연구하고 이론을 검증하기 위한 목적으로 출발하여 2015년부터 구축된 데이터베이스(seshatdatabank.info/)를 인터넷을 통해 제공하고 있다. 이 데이터베이스는 전 세계를 30개의 주요 권역으로 나누고 지난 1만 년 이래 존재했던 414개의 사회의 역사적 정보를 담고 있다. 데이터는 인류 사회의 특징을 9개의 핵심 속성, 예를 들어 사회 규모, 경제, 정부형태, 통화 체계, 종교 등으로 나누고 세부적으로는 51개의 변수로 각 사회의 정보를 표현한다. 각각의 속성과 변수는 시간 축을 담고 있어 특정 시대에 있었던 사회의 변화양상을 추적할 수 있다. 또한 데이터베이스는 각각의 서술 표현과 정보마다 어떤 연구에서 가져온 것인지 출처를 명기하고 있어 신뢰성을 확보하고 있으며, 이를 이용해 연구에 활용하기 쉬운 데이터 구조를 지니고 있다. 세스헤트는 첫 공개 이후로 수많은 연구에 활용되고 있으며 다양한 학회와 책을 통해 소개되고 있는 현재진행형인 프로젝트이다.

논문의 기술적인 내용은 데이터과학에서 많이 쓰이는 '주성분분석(PCA : Principal component analysis)' 알고리즘을 중심으로 다루고 있다.[15] 주성분분석 기법은 여러 개의 변수로 이루어진 데이터를 단순한 차원으로 압축해서 표현하는 것으로, 이 연구에서는 사회를 기술하는 51개의 변수 중 수량화가 가능한 핵심적인 부분만을 뽑아 9개의 특성 변수로 414개의 사회의 특성을 수량화 한 다음, 이를 분석해 몇 개의 중심축으로 여러 사회들을 분류할 수 있는지 살펴보고 있다. 〈그림 3-A〉는 이러한 특성과 변수의 구성과 관계를 설명하고 있으며, 〈그림 3-B〉는 주성분분석 결과를 해석한 그래프로, 데이터를 9개의 중심축(주성분)으로 나눈 결과, 놀랍게도 단 하나의 중심축(PC1)만으로도 사회복합성의 변화와 진화 양상을 의미 있게 나타낼 수 있음을 보이고 있다. 이는 사회의 발전 정도를 단 하나의 숫자, 즉 지표로 표현할 수 있다는 의미이다. 연구자들은 이 중심축을 사회복합성을 표현하는 단일한 지표로 보고, 시간 흐름에 따라 30개 권역의 사회 각각의 지표 값이 어떤 양상으로 변화하는지를 검토한다(〈그림 4〉). 예로 들어 이탈리아 지역의 정치체들은 로마 공

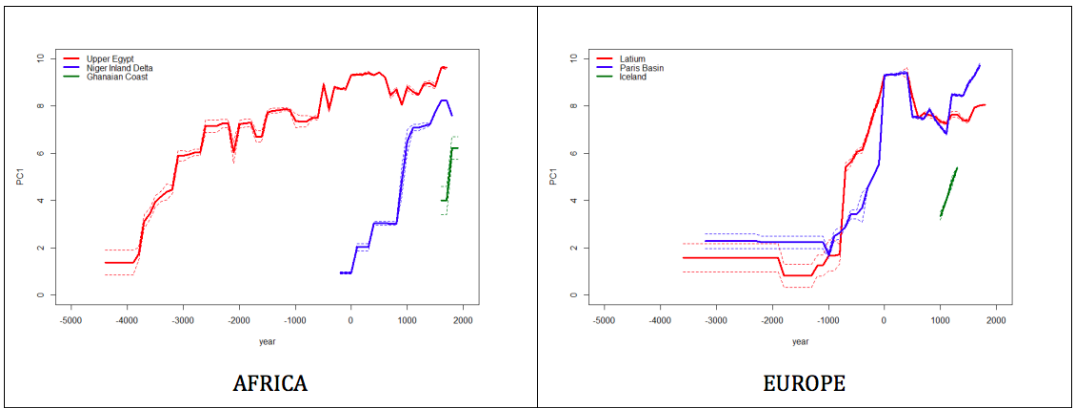
13) Turchin, Peter · Currie, Thomas E. · Whitehouse, Harvey · François, Pieter · Feeney, Kevin · Mullins, Daniel · Hoyer, Daniel, et al., "Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization.", *Proceedings of the National Academy of Sciences* 115: 2, 2018, pp. 144~151.

14) Turchin, Peter · Whitehouse, Harvey · François, Pieter · Hoyer, Daniel · Alves, Abel · Baines, John · Baker, David, et al., "An Introduction to Seshat: Global History Databank.", *Journal of Cognitive Historiography*, 2019.

15) 주성분분석(PCA : Principal component analysis)은 데이터를 여러 차원, 즉 벡터라는 형태로 표현할 때, 데이터가 어떻게 분포되어 있는지를 파악하여 중복되거나 다른 차원으로 표현할 수 있는 경우, 차원을 줄이거나 변환하는 기법이다. A2 논문에서는 세계 각지의 사회가 9개의 차원으로 데이터를 표현하고 있던 것을 주성분분석을 통해 단 하나의 차원만으로도 경향성을 의미 있게 보일 수 있음을 나타내고 있다. 하나의 차원으로 사회를 표현할 수 있으면 주가처럼 그래프 하나로 그 사회의 시간에 따른 변화를 표현할 수 있다. 물론 이렇게 압축한 차원이 역사적으로는 어떤 의미를 담는지는 역사학의 해석이 필요하다. 이 논문은 과거 역사적 사건과 발전 정도가 이렇게 하나로 줄인 차원으로도 해석할 수 있는 가능성이 있다는 것을 별도 자료를 통해 제시하고 있다.



〈그림 3〉 (A) Seshat 데이터의 51개 속성을 9개의 복잡성 특성(complexity characteristics : CC)으로 나누고 비슷한 것들을 묶은 관계도. 각 CC간의 연관 관계는 연결선의 두께와 색상에 비례한다.
(B) 주성분분석의 결과(가로축)의 분산 기여도(세로축) 비율 : 첫 번째 주성분인 PC1 하나만으로도 9개 CC의 특성을 80% 정도 드러낼 수 있다.
(C) 9개 CC(가로축) 각각을 B에서 찾은 단일 주성분 PC1에 비교할 때의 일치도. 9개 특성 모두를 비슷한 정도로 PC1 하나로 표현할 수 있다.(출처 : 논문 A2의 Figure 2.)



〈그림 4〉 10개 지역 중 아프리카와 유럽의 중요 지역의 PC1 주성분의 궤적. 가로축은 시간 흐름, 세로축은 같은 축척으로 바꾼 PC1 주성분의 값. 시간에 따라 점차 늘어나는 경향을 볼 수 있다.
(출처 : 논문 A2 보충자료의 Figure SI6. 일부)

화정과 제국 시기에 사회복합성이 극적으로 상승하여 최고 수준에 이르렀다가, 로마 이후는 조금 하강한 상태를 지속적으로 유지하고 있는 것을 볼 수 있다(〈그림 4〉 우측 붉은색 그래프). 연구자들은 이러한 사회의

복합성 변화가 대체로 역사적 사건과 비교하여 해석 가능한 여지는 충분하나 성급한 일반화는 주의해야 함을 설명한다. 논문은 사회복합성으로 표현되는 인류 사회의 여러 특성은 시간이 지날수록 누적되어 계속 증가하는 경향이 공통적으로 관찰된다고 결론짓고 있다.

인류 사회들의 역사 궤적과 이를 데이터로 종합하여 분석한 결과가 많은 부분에서 일치한다는 사실은, 곧 세스헤트 데이터의 품질이 높은 수준이라는 점을 간접적으로 입증하고 있으며 저자들 또한 그 부분을 드러내고자 한 것으로 보인다. 이는 세스헤트의 가장 큰 장점과도 연결되는데, 데이터베이스의 내용 각각에 대해 리뷰와 업데이트 요청이 가능하게 만든 그 열린 구조에서 기인한다. 처음부터 각 분야의 전문가들이 신뢰도 높은 데이터를 구축했고, 이후에도 지속적으로 그 품질을 유지하고 최신 연구 내용을 업데이트하면서 선순환 구조에 들어서고 있다.[16] 최근 『네이처』에서도 세스헤트를 이용한 연구결과가 실리면서 그러한 흐름은 가속화되는 것을 볼 수 있는데,[17] 앞으로도 높은 신뢰성을 지닌 훌륭한 역사 데이터베이스로 자리 잡을 것으로 보인다.[18]

(A3) 「신석기시대와 청동기시대 중국 고고학 유적지의 시공간 분포 패턴: 개요」[19]

이 연구는 중국의 신석기시대부터 청동기시대에 이르는 고고학 유적지를 데이터베이스화하고 데이터에서 중국의 고대 지역 문화가 시간 변화에 따라 어떤 양상을 보이는지 고찰하고 있다. 가장 중요하게 쓰인 자료는 『중국문물지도집(中国文物地图集, Atlas of Chinese Cultural Relics, ACCR)』으로, 이는 중국의 고고학 유적 발굴 결과를 중앙의 국가문물국(國家文物局)에 모두 등록하게 만든 뒤 그 결과를 종합해 출간한 자료집이다. 논문이 작성된 시점까지 발간된 총 25권의 자료집을 디지털화하여 초기 신석기시대부터 초기 철기시대에 이르는 51,074개 유적지 데이터베이스를 구축하였다. 이 데이터베이스에는 중국 본토 대부분의 유적지 정보가 들어있는데, 선행연구(Wagner et al. 2013)에서 진행했던 ACCR 11권에 대한 데이터베이스에 새로운 결과 – 주로 서부와 남부 지역 – 를 추가한 것이다. 이 연구에서는 단순히 기존 자료를 디지털화만 한 것이 아니라 유적지의 위치 정보를 지리정보시스템(GIS)을 이용하여 특정하고 있는데, 이를 통해 시간과 공간축 모두를 이용하여 데이터를 보다 높은 차원으로 분석할 수 있다. 이 두 개의 축을 모두 이용한 시각화 예시는

16) A2의 저자를 살펴보면 사회학, 역사학, 고고학, 환경학, 컴퓨터공학, 경제학, 데이터과학 등 다방면의 연구자를 볼 수 있으며, 세스헤트 운영진과 연구진 명단에서도 이러한 다양성을 확인할 수 있다. 심지어는 데이터분석 전문 기업도 포함되어 있어 이상적인 산학 연구의 사례로 간주할 수도 있겠다.
(세스헤트 운영진 정보: http://seshatdatabank.info/seshat-about-us/seshat-who-we-are/ 검색일: 2020.02.12.)

17) Whitehouse, Harvey · François, Pieter · Savage, Patrick E. · Currie, Thomas E. · Feeney, Kevin C. · Cioni, Enrico · Purcell, Rosalind, et al., "Complex societies precede moralizing gods throughout world history.", *Nature* 568, 2019, p. 226.

18) 다양한 분야의 연구자들이 직접 세스헤트를 이용하여 프로젝트를 수행하고 연구 결과를 여러 학회와 저널에 소개하고 있다. 홈페이지에서 많은 활동이 매년 이어지고 있음을 확인할 수 있다.
(참고: http://seshatdatabank.info/publications/ 검색일: 2020.01.15.)

19) Hosner, Dominic · Wagner, Mayke · Tarasov, Pavel E. · Chen, Xiaocheng · Leipe, Christian, "Spatiotemporal distribution patterns of archaeological sites in China during the Neolithic and Bronze Age: An overview.", *The Holocene* 26: 10, 2016, pp. 1576~1593.

〈그림 5〉에서 확인할 수 있다. 〈그림 5〉는 시간을 가로축으로 두고 고대 중국의 유적지 수를 세로축으로 잡아 시대별 유적 변화 양상을 省 단위로 볼 수 있는 그래프로, 해당 기간이 어떠한 문화인지를 같이 표시하여 신석기부터 철기에 이르기까지 어떤 특징을 보이는지 한눈에 파악할 수 있다. 논문에서는 이외에도 데이터의 특성을 다양한 차트와 그래프, 지도로 나타내고 있는데, 기존의 중국 고고학과 선사시대 연구 결과를 효과적으로 뒷받침하는데 활용될 수 있음을 주장하고 있다. 그러한 주장은 이어진 후속 연구로 입증되었는데, 이 논문을 인용한 연구 중 2건이 『네이처』에, 한 건은 『사이언스』 자매지에 실렸고 그중 두 개 논문에서는 핵심 연구 데이터로 이용되었다.

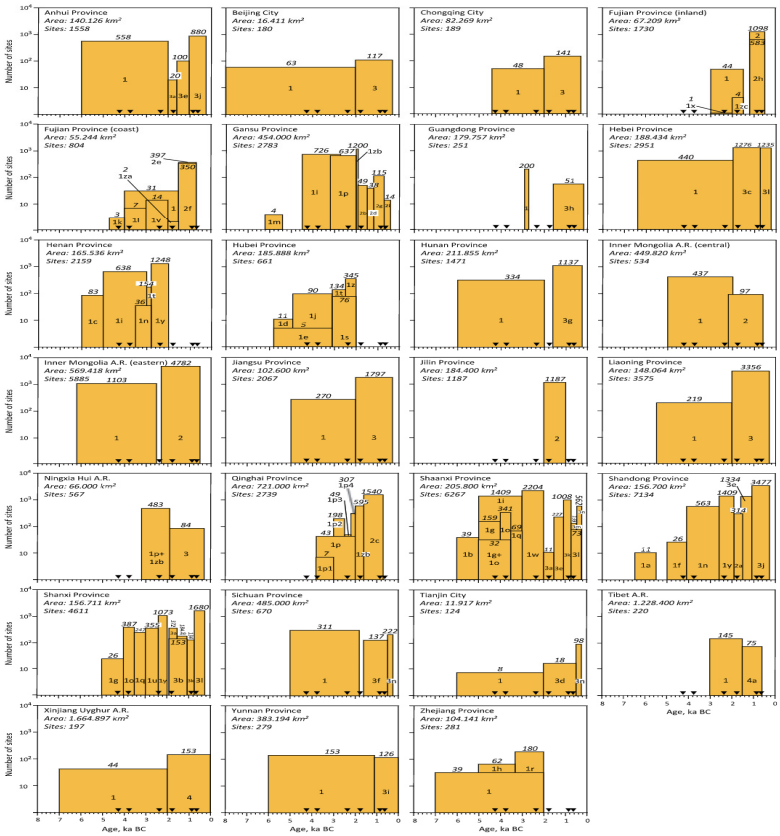


Figure 3. Quantitative changes in the number of archaeological sites (numbers in italic) across the analysed provinces and regions along with the cultural and chronological sequences. For the numbered culture names see Figure 2 caption. Black triangles indicate representative time slices chosen for mapping (Figure 4).

〈그림 5〉 연대순으로 분석된 중국 전역의 유적지 수의 정량적 변화 그래프. 가로축은 기원전 8,000년부터 0년까지 시간 정보를 담고 있고, 세로축은 유적지 수를 로그(log) 단위로 표시한다(이탤릭체 숫자는 유적지 수). 상자 안의 숫자는 시대와 문화의 약자이다(1: 신석기 문화, 2: 청동기 문화, 3: 하상주 시기, 4: 철기를 포함한 문화). 숫자 뒤의 알파벳은 그 시기와 겹치는 중국의 여러 문화를 의미한다(예: 1g – 초기 양샤오 문화, 3b – 구분이 어려운 하상대 문화 등) 논문에는 27개 省 그래프 전체가 있으나 여기서는 일부만 수록했다.
(출처 : 논문 A3의 Figure 3. 일부)

세 논문에 나타나는 주요 경향을 종합하자면, 첫 번째로 역사 데이터 구축은 그 자체로 쉽지 않은 일이라는 것이다. 서로 다른 데이터를 통합하는 과정은 특히 주의해야 하는 것으로, 엄밀한 기준하에 수행되지 않으면 본래의 의미를 상실할 수도 있다. 이렇게 통합된 데이터는 역사적 맥락과 사실에 부합하는지를 검증해야만 후속 연구에서도 의미 있게 활용될 수 있으므로 구축 과정에서 엄격한 검증은 필수이다. 두 번째는 최근 학계에서도 제대로 만든 데이터의 가치를 높이 평가하기 시작했다는 사실이다. 이는 앞서 소개한『사이언티픽 데이터』같은 전문 학술지의 등장으로도 확인할 수 있는 사항이지만, 이런 데이터만 모아 전문적으로 공유하고 제공하는 플랫폼이 다양하게 나오고 있으며,[20] 구축된 데이터들이 연구에 지속적으로 활용되고 권위 있는 학술지에 잇달아 실리고 있다는 점 또한 그러한 경향을 입증해준다. 특히 A3 논문처럼 발굴 자료와 보고서가 분석 가능한 형태의 공개 데이터베이스로 제공된다면 고대 문명 연구에 유용하게 사용될 수 있을 것이다. 이처럼 잘 구축된 데이터가 있다면 다양한 방식의 분석과 해석이 가능하다는 것을 다음 장에서 살펴보도록 한다.

Ⅲ. 데이터 분석: 자연스럽게 스며든 데이터과학

디지털 또는 컴퓨터를 활용한 연구방법론은 학문 분야별로 주로 사용되는 대표 기술이 있다. 인문사회학을 먼저 살펴보자면, 문학의 자연어 분석, 사회학의 네트워크 분석 기법이 대표적인 '학과－연구방법론' 짝꿍이다. 자연과학과 공학에서 보자면 지리학은 지리정보시스템(GIS) 분석, 물리학은 통계물리와 복잡계연구, 컴퓨터공학에서는 최근 떠오르는 기계학습과 인공지능이 학과－연구방법론 짝으로 연결되어 있다. 그리고 통계학에 뿌리를 두는 데이터과학은 앞서 이야기한 여러 기법을 포괄적으로 다루며 데이터 자체의 완결성 있는 분석과 체계화, 시각화에 중점을 둔다. 고고학은 일찍이 디지털 기술・데이터 분석을 다방면으로 적용하고 활발하고 이용하는 학문 분야로, 앞서 언급한 학문 분야별 대표 연구방법론을 대부분 적용하고 있다. 이에 최신의 고고학 연구 사례 중 데이터 분석 측면에서 주목할 만한 논문을 선정하여, 데이터과학 방법론을 구체적으로 어떻게 적용하였는지 살펴보고자 한다.

이를 위해 '데이터 분석의 표준 프로세스'가 무엇인지를 정의하는 것이 필요하다. 데이터과학에서 정형화된 '표준'분석 프로세스는 쓰임에 따라 다르게 정의되지만, 공통적으로 언급되는 단계는 다음과 같다: 1) 기획 ⇒ 2) 수집 ⇒ 3) 정제 ⇒ 4) 분석(탐색, 모델링) ⇒ 5) 시각화 및 스토리.

분석 방향과 어떤 부분에 초점을 두느냐에 따라 위의 각 단계는 세분화되기도 하고 합쳐지기도 한다. 본 글에서는 1・2단계를 합쳐 '데이터 선정과 수집'으로, 3단계는 여러 자료의 통합을 중요한 기술로 보아 '데이

20) Kohler, Timothy A.・Buckland, Philip I.・Kintigh, K. W.・Bocinsky, R. K.・Brin, A.・Gillreath-Brown, A.・Ludäscher, B.・McPhillips, T. M.・Opitz, R.・Terstriep. J., "Paleodata for and from archaeology.", *PAGES Magazine* 26: 2, 2018, pp. 68~69.

터 정제 및 통합'으로 명명한다. 그리고 4・5단계는 명확히 분리되기보다는 섞여서 수행하는 경우가 많으므로 '데이터 탐색' 과 '다각도로 분석하기(분석 및 시각화)'로 나누기로 한다.

〈표 1〉 데이터 분석 프로세스

P1) 데이터 선정과 수집	연구에 사용할 수 있으며 가치가 있는지 그 배경을 확인한 뒤 검증 가능한 방식으로 데이터를 모은다.
P2) 데이터 정제 및 통합	불필요한 정보나 잘못된 부분을 찾아내어 제거하고, 정확한 데이터로 사용 가능하게끔 만드는 과정이다. 또한 출처가 다른 여러 데이터를 통합하는 경우가 많은데, 예를 들어 역사 자료와 지리 정보와 결합하는 등의 작업등이 해당된다. 또한, 이 과정은 해당 분야를 제대로 이해하는 연구자가 꼭 필요한 단계로 합리적이며 엄밀한 기준으로 정제하고 통합하는 것이 중요하다.
P3) 데이터 탐색	데이터 자체의 성격을 확인하는 단계로, 분포와 결측치 같은 전반적인 통계 지표를 만들어 특성을 확인한다. 이후 어떤 분석 방법을 사용할 것인지를 결정하려면 이 단계에서 특성을 파악해야 한다.
P4) 다각도로 분석하기 (분석 및 시각화)	이전에 알기 어려웠던 것을 찾고 가설을 뒷받침하거나 직관을 잘 드러낼 수 있는 부분을 설명하는 부분으로, 다양한 차트, 지도, 그래프 등을 이용해 이해를 높일 수 있다. 데이터과학의 여러 가지 분석 기법과 알고리즘이 이 상황에 맞는다는 것을 보이고 적용한 결과를 정리하는 것도 이 단계에서 많이 활용하는 방법이다.

〈표 1〉의 구분(P1・P2・P3・P4)은 다음에 소개할 논문에서 데이터과학의 프로세스 사용례를 효과적으로 표현하기 위해 본고에서 정의한 단계이지만, 기본적인 데이터과학의 작업 순서와 일치함을 다시 한 번 강조하고자 한다. 살펴볼 논문(A4・A5・A6)은『사이언스 어드밴스』에 최근 5년간 발표된 연구로,[21] 이번 장에서는 이 논문들이 데이터과학의 분석 프로세스를 빠짐없이 내포하고 있다는 것을 확인할 것이다. 각 과정마다 분석 프로세스의 어떤 부분이 대응되는지는 앞서 표기한 P1・P2・P3・P4로 기술하도록 하겠다.

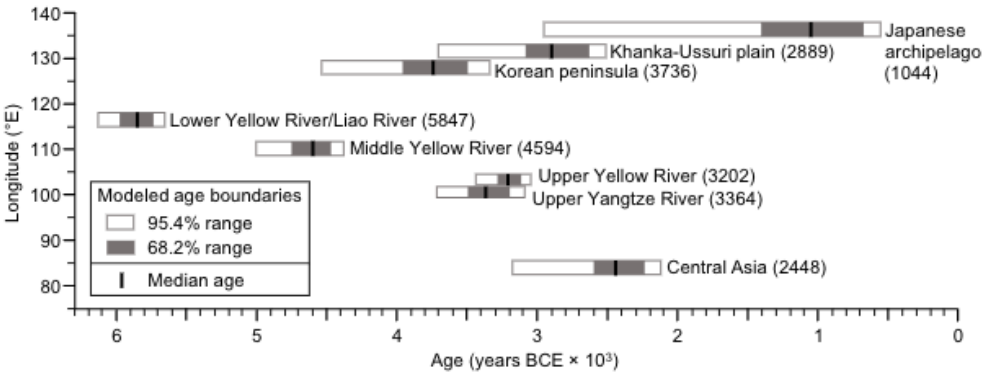
(A4)「동아시아 수수 농사의 불연속적 확산과 선사시대 인구 동역학」[22]

이 논문은 동아시아에서 일찍이 재배된 기장과 조가 지역적으로 볼 때 불연속적으로 확장된 경향을 분석한 것으로, 중국과 한국, 일본에서 수수(기장, 조를 모두 포함) 재배시기를 탄소동위원소법으로 추정한 170개의 고고학 유적 데이터를 이용한다(P1). 연구 범위를 8개의 지역 분류로 나누어 시기별 확산 경향을 알아보려고 하나, 각 지역 안에서도 유적지별로 재배 시점의 범위가 넓고 모수가 적어 신뢰도 있는 값을 사용하기 어려운 문제가 있다(P3). 이에 베이즈 모형(Bayesian Model)을 사용하여 지역별로 수수 재배 시점을 통계적으로 의미 있게 재구축할 수 있는데,[23] 그 결과는 〈그림 6〉에서 확인할 수 있다(P4). 아울러 지역별 수

21) *Science Advances*는 *Science*지에서 2015년 창간한 자매지로 오픈엑세스 과학 저널이다. 과학 전반을 모두 다루며, 사회과학과 컴퓨터과학 논문도 게재 가능하다.

22) Leipe, C.・Long, T.・Sergusheva, E. A.・Wagner, M.・Tarasov, P. E., "Discontinuous spread of millet agriculture in eastern Asia and prehistoric population dynamics.", *Science Advances* 5: 9, 2019.

수 재배 확산을 설명할 수 있도록 러시아-중국 경계 지역 고고학 보고서와 같은 최신 데이터도 동일한 형태로 만들어 통합한다(P2). 그리하여 동아시아와 중앙아시아 전역을 아우르는 수수 재배의 시간–공간 패턴을 만들어 지도와 시간축 그래프로 각각을 설명한다(P4). 수수 재배 확장 패턴이 실제 인구 패턴과도 일치하는지 교차 검증하기 위해 중국 전역의 고고학 데이터베이스(앞 장에서 소개한 A3 논문 데이터)를 이용해 총 51,074개 유적지 중 수수 재배 지역과 일치하는 40,696개 유적지 데이터를 뽑아(P1) 재배 시기 데이터와 통합한다(P2). 이후 유적지 숫자를 인구 분포로 가정하여 시간에 따른 인구 증가와 유적지 분포가 일치하는지 핵밀도추정(Kernel Density Estimate : KDE) 모형[24]을 이용해 비교한다(P4). 이를 이용해 중국 북중부 지역과 북동지역 등 여러 지역에서 인구 변화와 수수 재배 경향성이 일치하는 정도가 입증 가능한 수준임을 보이고, 더 나아가 분석 결과에 기반하여 일본 북부는 연해주를 거쳐 전래되었고, 중앙아시아는 중국 중부의 인구 이주를 통해 수수 재배가 확산되었을 것이라는 가설을 제기한다(P4).

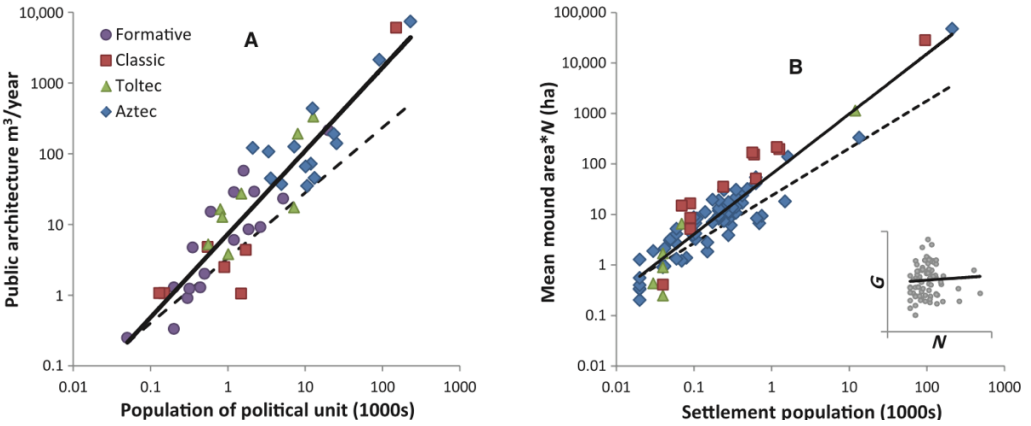


〈그림 6〉 베이즈 모형으로 동아시아 8개 지역의 수수 재배 증거를 담은 고고학 유적지 데이터(170개)의 지역별 재배시점의 평균값과 신뢰구간을 구한 결과. 각 지역의 평균 재배 시점은 괄호 안의 숫자를 참고할 것. 가로축의 눈금은 100년 단위로 기원전 6,000년부터 시작하고, 세로축은 경도를 나타낸다. 이 결과는 OxCal 이라는 탄소동위원소 데이터 분석용 소프트웨어로 구한 결과이다. 소스 코드는 논문의 추가 자료에 공개되어 있다.
(출처 : 논문 A4의 Figure 2.)

23) 베이즈 모형(Bayesian model)은 두 확률변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 베이즈 이론(Bayes theorem)을 기반으로 한 것으로, 이 경우는 데이터를 모형을 통해 전체 모수를 추론하는 데 사용한다. 즉 적은 수의 수수 재배 데이터를 그대로 이용하면 재배 시기의 오차 또한 그대로 활용하게 되므로 모수를 계산하여 충분한 데이터(모수)가 있었다는 것을 가정했을 때 재배시점이 어떻게 되는지 추론한다. 이때 재배시점은 확률로 표현하게 되므로 〈그림 6〉처럼 기간으로 나타나게 된다.

24) 핵밀도 추정 모형(Kernel Density Estimate)은 실제 변수가 가지고 있는 본질적인 특성을 파악하기 위한 방법이다. 여러 데이터는 특정 변수의 일면에 불과한데, 관측된 데이터를 이용해 거꾸로 그 변수의 특성, 여기서는 확률분포라 불리는 특성을 알아내고자 하는 방법이다. 이 확률분포가 어떠한 모양을 가지고 있을지를 알기 위해 밀도 추정 모형을 쓰는 것이고, 그 중 하나로 커널함수(Kernel function)을 이용하는 것이 핵밀도추정(KDE)이다. A4 연구에서는 유적지 수와 분포를 인구 증가 분포와 비슷하게 볼 수 있을지를 추정하기 위해 사용하였다. 즉 이 방법으로 검증되면 유적지 수와 인구수의 관계가 통계적으로 유의미하다고 볼 수 있다.

(A5) 「고대 사회의 취락 측정과 증가하는 반대급부」[25]



〈그림 7〉 (A) 정치체의 인구 규모 증가(가로축)과 공공건축물의 크기(부피)를 세로축에 놓은 그래프. 각 점은 멕시코시티에 위치했던 문명의 각 사회집단의 인구/건축물 크기를 나타낸다. 점선은 선형 증가일 때를 의미한다(가로-세로 변화량이 같은 경우). (B) 가로축을 정주인구, 세로축을 정주지의 평균 면적으로 본 그래프. 두 그래프 모두 시대와 상관없이 문명 사회들의 사회 발전도는 일정한 패턴이 있음을 보이고 있다(굵은선이 추세선). 그리고 그 패턴은 선형이 아닌 초선형 형태를 띠고 있다.(출처 : 논문 A5의 Figure 2.)

앞서 살펴본 연구(A4)가 농학과 고고학의 합작이라면, 이번 연구는 인류학과 물리학, 그중에서도 복잡계 네트워크 이론의 만남으로 볼 수 있다. 이 연구는 멕시코시티가 위치한 멕시코 분지(Basin of Mexico : BOM)의 선(先)스페인 시기 고고학 자료를 이용해 인구 증가와 건축물 규모 간의 관계를 패턴으로 일반화할 수 있는지를 고찰한다. 논문은 선행 연구(Ortman et al. 2014)에 기초해 가설을 세우고 그에 필요한 데이터와 그 형태를 먼저 정의하는데, 이미 기존 연구에서 활용한 방식과 비슷하게 데이터를 활용할 수 있기에 가능한 작업이다(P1 · P3). 데이터 수집은 1960년부터 1975년 사이 멕시코시티 고고학 발굴 자료를 집성한 전자 자료에서 시작하는데, 이 과정에서 가설을 입증하기 부족한 부분을 보강하기 위해 정주 패턴에 대한 연구 자료와 공공 墓葬 자료, 최신의 유적 분석보고서 등을 추가해 총 4,000여개의 유적 데이터를 구축한다(P1 · P2). 이후 시대별로 유적지의 면적과 인구 관계를 통계 분석한 결과를 이용해(P4), 가설로 세운 인구-면적 네트워크 공식의 계수를 측정하고, 이 값으로 사회구조에 따라 인구 분포와 수치 값을 합리적으로 예상할 수 있음을 보인다. 또한 시대와 상관없이 '정치단위체 인구와 공공 건축의 규모'와 '정주인구수와 거주 구릉지 면적'의 상관관계는 물리학에서 말하는 초선형 스케일링(Superlinear scaling)의 특징을 따르는 것을 통계량과 그래프를 이용해 밝히고 있다(P4)(〈그림 7〉 참고).[26] 여기서 가설로 세운 '적정 확장성(scaling)' 공식의 예

25) Ortman, Scott G. · Cabaniss, Andrew HF · Sturm, Jennie O. · Bettencourt, Luís MA, "Settlement scaling and increasing returns in an ancient society.", *Science Advances* 1: 1, 2015.

26) 초선형 스케일링(Superlinear scaling)이란 x, y축이 log로 측정되는 그래프에서 기울기가 1.0(=linear) 이상인 그래프를 의미한다. 예를 들어 기울기가 1.2인 초선형 척도에서는 x가 100% 증가할 때, y는 120% 증가한다는 의미이다. A5 논문에

측값보다 고고학 자료로 추정한 값이 경향은 일치하나 약간 상회한다는 점을 밝히고 있다. 이 연구는 인간의 정주－확장 이론이 현대 사회뿐만 아니라 고대 사회에도 적용될 수 있는 일반론이라는 것을 입증하기 위한 것으로, 고고학 자료를 이용한 전형적인 사회과학 연구의 형태를 보여준다.

(A6) 「환경 변화가 아시아 전역의 농업 혁신과 교환을 자극했다」[27]

이 논문은 기후 데이터와 고고학 기록을 이용해 10,000여 년 전부터 유라시아의 기후 변화가 농업을 어떻게 다변화시키고 다양한 형태의 교역과 적응을 이끌어냈는지 고찰한다. 먼저 홀로세 시기의 기후 변화를 알기 위해 이 기간의 기후 변동 모형을 다룬 선행 연구데이터와 '지구기후 네트워크(Global Historical Climatology Network : GHCN)' 데이터베이스[28]에서 최근 몇십 년간의 범지구적 기후 데이터를 모아 통합한다(P1 · P2). 그리고 각 곡물의 생육 가능 온도 범위, 즉 기후 생태적 지위를 확률로 표현한 논문 저자의 선행연구를 이용하여 6개의 주요 곡물(밀, 보리, 메밀, 조, 기장, 벼)의 분포 범위를 지도와 그래프로 나타내고 있다(P3 · P4). 이러한 기후 모형이 실제 고고학적 기록과 일치하는지 확인하고자 아시아 전역의 곡물 고고학 데이터를 여러 논문과 보고서에서 직접 가져와 통합하여 활용한다(P1 · P2). 이렇게 고고학적 데이터와 기후 변화 데이터가 시간축과 공간축으로 구축된 경우, 여러 차원으로 축을 바꿔가며 다각적인 관점으로 접근할 수 있다. 이 연구에서는 이를 통해 얻은 유의미한 결과를 다양한 그래프로 여러 페이지에 걸쳐 보이고 있다(P4, 이후 설명도 모두 같은 단계). 유적지 별로 곡물들의 생태적 지위 분포가 시간 흐름에 따라 어떻게 변화하는지를 담은 그래프(〈그림 8〉)만으로도 농업 작물의 다변화가 특정 시점부터 보편적으로 발생한 현상임을 쉽게 파악할 수 있다. 그리고 농업에 관한 최근의 고고학 연구와 해석을 이 데이터 분석이 뒷받침할 수 있는지를 설명하고 있는데, 오늘날 중국의 곡물 재배 현황이 과거의 기록과 어떤 연관성이 있는지를 설명하기 위해 추가 데이터를 가져와 비교하기도 한다(P4). 이 연구는 결과 데이터뿐만 아니라 논문에 설명된 모든 그래프와 표, 그리고 최종 데이터 전체를 재현할 수 있는 프로그래밍 코드도 공개하여 투명성과 후속 연구의 가능성을 높였다는 점에 주목할 만하다.[29]

서는 인구수를 x 축으로 잡았을 때, 건축물의 크기와 규모를 y축으로 보면, 아즈텍 문명 등 멕시코시티에 있던 문명들이 공통적으로 인구 증가에 비해 건축물의 규모가 더 커졌다는 점을 보이고 있다.

27) Guedes, Jade d'Alpoim · Bocinsky, R. Kyle, "Climate change stimulated agricultural innovation and exchange across Asia.", *Science Advances* 4: 10, 2018.

28) 지구기후 네트워크(Global Historical Climatology Network : GHCN)는 미국 국립기후센터에서 관리하는 기온, 강수량, 기압 데이터베이스로 15,000개가 넘는 고정 관측소의 데이터를 수집하여 저장한다. 가장 오래된 기록은 1701년부터 수집된 것으로, 이를 통해 과거 200년 전부터 존재하는 기후 데이터를 이용하여 지구 기후가 어떻게 변화하였는지를 알 수 있다.

29) 데이터 구축부터 분석, 그래프 생성까지 연구 결과로 발표된 모든 과정을 수행할 수 있는 프로그램 소스코드가 온라인 소스코드 공개플랫폼인 깃허브(Github)에 공개되어 있다.
(A6 논문의 소스 코드: https://github.com/bocinsky/guedesbocinsky2018, 검색일: 2020.01.19.) 데이터 분석이 핵심인 연구는 이처럼 연구 데이터뿐만 아니라 코드까지 공유하는 추세이다.

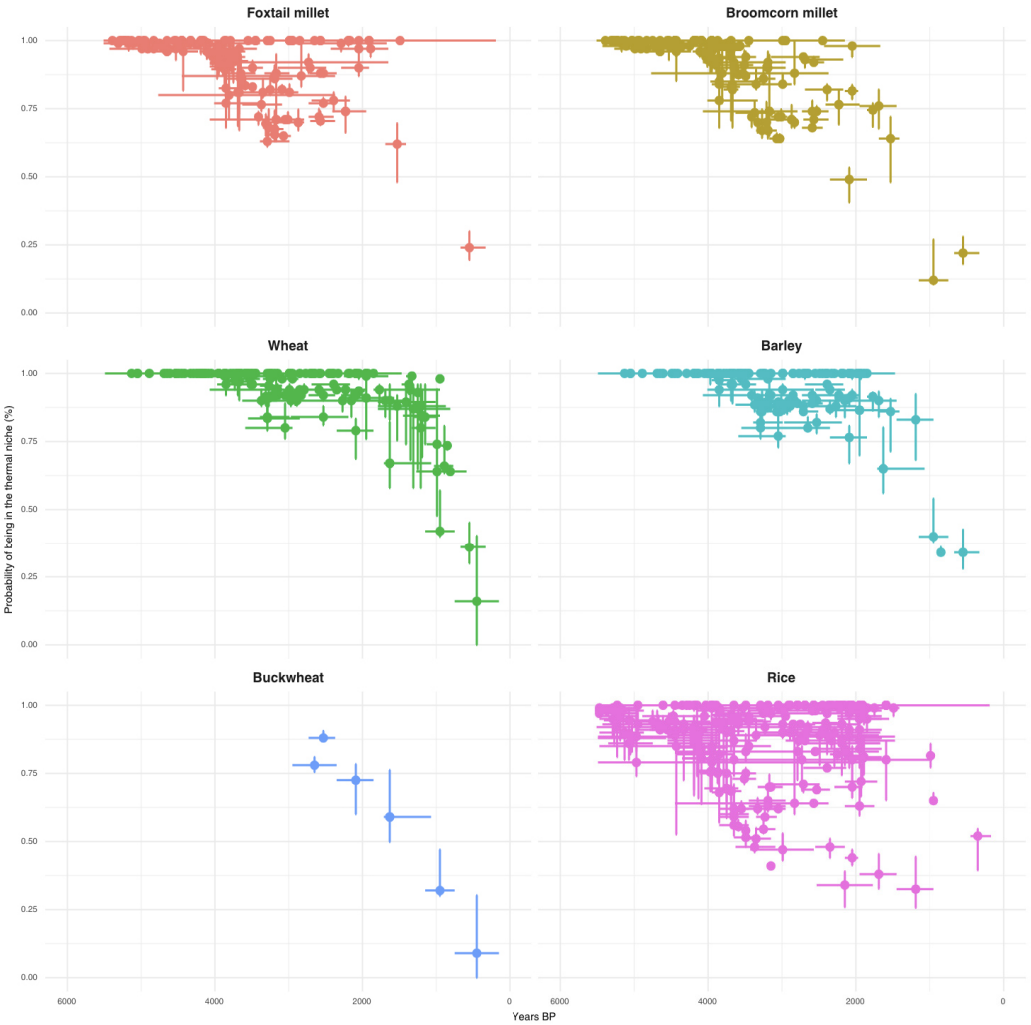


Fig. 3. Calibrated radiocarbon ranges for each site in our database (*x* axis) and the probability of each of these sites being in the niche during that same phase of occupation (*y* axis). An interactive version of this figure that labels each of the cross-plots and enables zooming is available as data file S2.

〈그림 8〉 (좌상단부터) 조, 기장, 밀, 보리, 메밀, 쌀의 생육 환경적 지위를 확률값으로 표현한 그래프. 가로축은 시간대(과거 6000년부터 현재까지 연도를 BP로 표현)를 나타내고, 세로축은 해당 곡물이 어느 정도의(온도)생태적 지위를 가지고 생육이 가능했는지 확률로 나타낸 값이다. 각 점은 곡물 재배 흔적이 발견된 유적지를 의미한다. 점의 가로선은 방사성탄소연대로 보정된 유적지의 연대 범위, 세로선은 생태적 지위의 확률값 범위이다. 이 그래프에서 알 수 있는 것은 과거(왼쪽)에는 단일한 곡물을 재배하는 경향을 보였다는 것으로, 이후 시간이 흐르면서 같은 유적지에서 여러 곡물 재배 흔적이 나오면서 이에 곡물 각각의 생태적 지위가 모두 낮아지는 것을 모습을 관찰할 수 있다. (한 지역에서 단일한 곡물만 재배하면 그 곡물은 시장을 독점하는 것처럼 상대적 생태적 지위가 높다. 한 지역에서 여러 곡물이 재배되면 서로 생존 경쟁이 일어나므로 생태적 지위, 즉 생존 확률이 줄어든다. 그래프의 오른쪽으로 갈수록(시간이 흐를수록) 곡물의 생존 확률(세로축)이 낮아지는 것을 확인할 수 있다.) (출처 : 논문 A6의 Figure 3.)

이 장에서 살펴본 논문들은 전개 방식과 여러 종류의 통계분석 수행 과정, 그리고 결과를 표현하는 다양한 그래프까지 데이터과학의 분석의 전형을 공통적으로 드러내고 있다. 또한, 고고학 발굴 데이터를 주요 자료로 사용하면서 서로 다른 데이터를 통합하고 엮을 수 있도록 하는 과정에 많은 공력을 쏟고 있다. 이는 고고학과 불가분의 사이인 고대 문명 연구에 참고할 여지가 많을 것으로 보인다. 무엇보다도 자료가 풍부하지 못한 고대의 특성상, 여러 데이터를 수집하고 통합하는 데이터과학의 방법론은 많은 도움을 줄 수 있다. 이렇게 만든 데이터와 프로그래밍 코드를 온라인 공유 플랫폼으로 배포하여 후속 연구가 쉽게 일어날 수 있게끔 한 것도 위 연구의 공통된 특징이다. 이제 이러한 데이터 분석을 넘어, 연구 대상과 그 주변 상황, 그리고 시간 흐름에 따른 변화 그 자체를 직접 재현하며 들여다볼 수 있는 방법론을 다음 장에서 확인해 보도록 하겠다.

Ⅳ. 데이터 기반 시뮬레이션: 마르지 않는 'What-If' 샘물

시뮬레이션이란 사전적으로는 '복잡한 문제나 현상을 해석하고 해결하기 위해 실제와 비슷한 모형을 만들어 행하는 모의실험'을 뜻한다. 오늘날 컴퓨터와 디지털기술의 발전으로 대부분 시뮬레이션이라 하면 컴퓨터를 으레 떠올리는데, 프로그램으로 작성된 모형을 실행하고 그 과정과 결과를 보는 것으로 단순화하여 이해해도 큰 무리는 없어 보인다.

이 시뮬레이션을 연구에 활용하는 것은 더 이상 자연과학과 공학의 전유물이 아니다. 이미 고고학에서는 '행위자 기반 모형 ABM(Agent-Based Models)'이란 방법론을 연구에 적극 활용하고 있다.[30] 이는 주로 고고학적 발굴 결과를 토대로 당시 사회와 인간 행동양식이 어떠했을지 파악하는데 유용하게 활용되고 있는 방법론이다. ABM은 환경과 행위 주체인 인간의 행동과 상호작용을 정의하고 프로그램화하여 시간의 흐름에 따라 어떤 패턴을 보이는지를 분석한 뒤, 그 결과가 고고학적 증거에 잘 들어맞거나 정황을 합리적으로 설명할 수 있는지 여부를 검토한다.

이처럼 시뮬레이션은 현상 예측과 법칙의 증명에 많이 활용되는데, 고고학이나 역사학처럼 당대를 그려보는데 필요한 지식의 공백이 많아 추론의 여지가 많을수록 도움이 된다. 알고 있는 부분은 고정 값으로 두고, 미지의 영역은 수시로 바꿔 넣어볼 수 있는 변수로 만들어 모형을 만든 뒤, 가상으로 시뮬레이션을 수백 수천만 번 실행해볼 수 있다. 그 중, 특정 패턴이 실제 역사적 사실과 고고학적 증거와 부합하면서 동시에 재현

30) 각 Agent(행위자)와 행위자의 행동을 미리 정의해두고, 찾고자 하는 변수를 입력한 뒤 실행하면 가상의 시간 흐름에 맞게 행위자들과 주변 환경이 어떻게 변하는지를 직접 관찰할 수 있다. 컴퓨터 게임을 ABM의 활용 사례로 바라보는 것도 이해하는데 도움이 된다. 고고학에서 ABM이 어떻게 사용되는지를 보려면 다음 글을 참고할 것:
Romanowska, Iza · Stefani Crabtree · Benjamin Davies · Kathryn Harris 2019 "Agent-based Modeling for Archaeologists. A step-by-step guide for using agent-based modeling in archaeological research (Part I of III)." *Advances in Archaeological Practice* 7, 2019, p. 178.

가능함과 정합성을 지닌다면, 그 패턴에 쓰인 변수 값을 바로 궁금했던 질문의 유력한 답으로 추정해 볼 수 있다. 더 나아가 다양한 합리적 모형을 별도로 만들고 각 시나리오별로 어떻게 진행되는지 시간 흐름에 따른 변화를 관찰할 수 있다. 즉 시뮬레이션 방법론은 'What-If'를 수없이 많이 시도해볼 수 있는 강력한 도구인 것이다. 또한 시뮬레이션 방법론은 유연함과 확장성이라는 이점도 있다. 자신의 연구가 시뮬레이션으로 표현 가능한 모형의 형태로 있다면 기존의 데이터를 뒤엎는 새로운 자료가 나왔을 때, 그 모형을 수정해서 새로운 데이터를 반영해 실행해보고 좀 더 나은 설명으로 모형을 진화시킬 수 있다. 그리고 새로운 연구 성과를 업데이트하면서 다음 버전의 시뮬레이션 모형으로 관리하는 형태로 연구를 지속적으로 확장할 수도 있을 것이다.

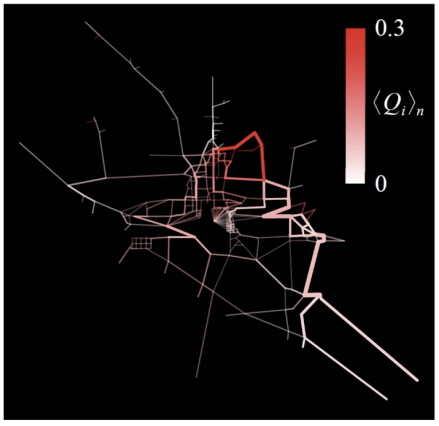
한 가지 흥미로운 사실은 이미 고고학에서는 20여 년 전부터 시뮬레이션을 이용한 연구 성과가 나오기 시작하였다는 점인데, 그 강력함에 비해 분과 이상으로 널리 쓰이는 것으로 보이지는 않는다. 몇 가지 원인을 추측해보건대 우선 시뮬레이션 결과를 보이는 방식이 그 제약점 중 하나로 보인다. 시뮬레이션을 활용한 연구를 논문 또는 책처럼 활자화된 방식으로 설명하다 보면 그 특유의 생명력과 재현력을 확인하는 데 한계가 있을 수밖에 없다. 이는 시뮬레이션 프로그램을 동작시켜 값을 직접 바꿔가며 변화하는 결과를 확인할 수 없기 때문이다. 전통적인 출판 기반의 연구 성과 공유는 다른 연구자들의 시뮬레이션 프로그램을 접하기 어렵게 하고, 이를 이용한 후속 연구가 활성화되기 어려운 환경을 만든다. 연구자들이 시뮬레이션 방법론을 충분히 인용하고 활용하려면 논문이라는 형태가 아닌 동작 가능한 형태로 공유되고 다시 사용할 수도 있게끔 제공될 필요가 있다. 일례로 다른 연구자들도 직접 시뮬레이션을 실행해볼 수 있도록 인터넷에 동적 웹사이트를 만들어 프로그램을 올리는 방법이 가능할 것이다. 그러나 IT 기술이 발달한 이 시점에도 이러한 사례는 드문 것으로 보이고, 그나마 있는 경우도 다른 연구자들이 이해할 정도의 직관적, 시각적인 요소가 떨어져 여전히 아는 사람만 계속 활용하게 되어 외연 확장이 쉽지 않은 현실이다.[31] 불행 중 다행이라면 오늘날 디지털 인문학의 발흥이 이러한 문제점을 차차 해결할 가능성이 있다는 점이다. 이를 위해 시뮬레이션 코드와 프로그램을 공유하는 수준을 넘어, 해당 분야를 잘 모르는 동료 연구자들에게도 바로 사용할 수 있도록 접근성을 높이는 것이 필요하다. 그리하면 시뮬레이션 방법론 또한 빠르게 외연을 확장하여 고고학의 주요 연구 방법론으로 자리 잡을 수 있을 것이다.

이에 이러한 시도가 가능한지 알아보고자 구체적인 사례를 살펴보기로 한다. 앞서 소개한 ABM은 아니지만 시뮬레이션 방법론을 사용한 최근의 연구 사례 중 『네이처』와 『사이언스 어드밴스』에 등재된 두 건(A7・A8)의 논문을 살펴보면서 시뮬레이션 기반 연구의 특징과 장점을 검토해보기로 한다.

31) CoMSES network(https://www.comses.net/) 라는 행위자기반모형(ABM) 연구자 커뮤니티가 잘 알려진 사례이다. 이곳에서는 여러 연구자들이 자신의 연구 결과를 논문과 시뮬레이션 모형을 같이 올려 공유한다(검색일: 2020.02.12.).

(A7) 「앙코르의 종말 : 기후 변화에 대한 도시 인프라의 체계적 취약성」[32]

중세 캄보디아의 크메르(Khmer)[33] 제국의 수도 앙코르(Ankor)가 쇠락한 이유로는 여러 가설이 나오고 있는데, 이 연구는 그중 하나인 기후 변화로 인한 수리 체계의 붕괴를 다루고 있다. 앙코르는 13세기 세계 최대의 도시였는데 약 1,000㎢에 달하는 도시 전역은 수많은 운하, 저수지, 제방, 해자 등으로 연결된 수로로 덮여 있었다. 이 수로 인프라는 600여 년 동안 앙코르를 홍수와 가뭄에서 효과적으로 벗어날 수 있도록 해 준 것으로 보인다. 최근의 고고학 증거는 앙코르가 여러 해 동안 수로 인프라의 손상을 입었고, 그로 인해 대규모의 도시 분절화가 나타났음을 입증한다. 이 논문은 홍수가 발생할 때 수로의 구조적 형태에 손상을 입을 가능성과, 그것이 누적되었을 때 결국 특정 물길이 퇴적되어 사용할 수 없게 되는 과정을 수식과 모형을 통해 시뮬레이션으로 보이고 있다. 순서를 보자면 홍수의 강도, 침식 한계값, 퇴적 경향성을 연구자가 임의로 설정할 수 있는 변수로 두고 지형과 상관없는 위상학적으로 단순화된 구조에서 동작하는 단계별 손상 모형을 먼저 설계한다. 이 모형이 시스템 붕괴를 관찰할 수 있는 합리적 설명인지 다양한 변수를 넣고 실험하여 적정 값을 예측한다. 그다음 단계는 실제 앙코르의 수로를 네트워크로 표현하는 것인데, 고고학 발굴 결과를 이용한 과거 지도를 바탕으로 수로의 길이, 너비 등을 수치화한다. 그런 후 앞서 만든 모형에 앙코르 수로 체계를 넣고 다양한 강도의 홍수를 상정하여 시뮬레이션하고 그 결과를 확인한다(〈그림 9〉 참고). 모형은 이러한 단계별 손상이 홍수의 강도에 따라 급격히 증가함을 입증하는데 고강도의 홍수가 연속적일 때 특히 시스템의 붕괴가 빠르게 올 수 있음을 정량적으로 설명한다. 논문은 14세기 말 동남아시아에 강우량이 특히 많았다는 고고학 증거 또한 이 시뮬레이션 결과가 당시 상황을 더욱 설득력 있게 설명하고 있음을 밝히고 있다.



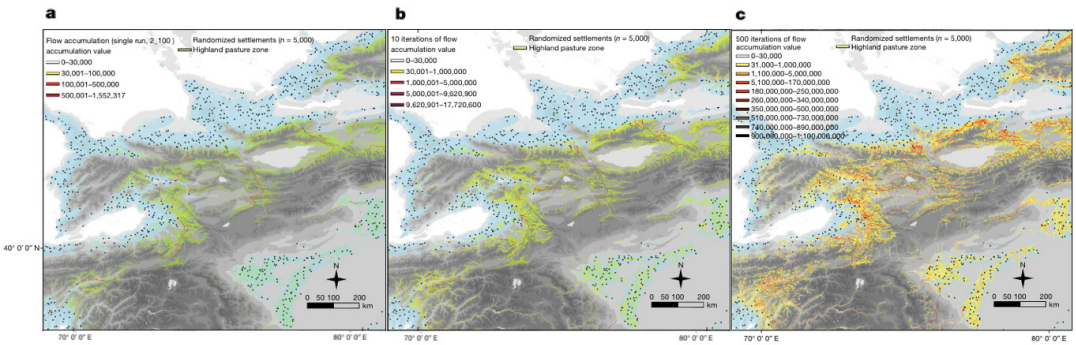
〈그림 9〉 앙코르의 수로 네트워크와 시뮬레이션 결과를 나타낸 지도. ⟨Qi⟩n은 수로의 손상 정도를 나타내는 함수로, 이 경우는 n=3,000 번 반복했을 때 각 수로(연결선)가 입은 손상이 어느 정도인지 색으로 표현했다. 붉은색이 진할수록 손상이 심한 것으로, 그 지역에 홍수가 일어날 가능성이 커진다. (출처 : 논문 A7의 Figure 5.)

32) Penny, Dan · Zachreson, Cameron · Fletcher, Roland · Lau, David · Lizier, Joseph T. · Fischer, Nicholas · Evans, Damian · Pottier, Christophe · Prokopenko, Mikhail, "The demise of Angkor: Systemic vulnerability of urban infrastructure to climatic variations.", *Science advances* 4: 10, 2018.

33) 크메르 제국(Khmer Empire)은 9세기부터 15세기까지 인도차이나 반도에 존재한 왕국으로, 현재 캄보디아의 원류가 된 나라이다. 앙코르 와트(Angkor Wat)와 앙코르 톰(Angkor Thom)과 같은 화려한 유적을 남겼다. 13세기 초에 가장 융성하였으나 이후 쇠락을 거듭해 타이의 아유타야(Ayuthaya)에 의해 멸망하고 이후 캄보디아 왕국으로 명맥을 이어나갔다.

(A8) 「유목 생태가 형성한 아시아 실크로드의 고지대 지리학」[34]

이 연구는 역사학계에서 대체로 합의된 결론인 실크로드 교역의 네트워크적 관점을 소개하면서, 그에 비해 오아시스나 도시가 없는 고지대 목축 지대에서 명확하지 않은 네트워크 연결성을 지적하고 있다. 이에 논문에서는 그 부분을 밝히고자 시뮬레이션을 활용하는데, 높이 750-4,000m 사이의 고지대에 위치한 유목 사회들이 계절성 목축 이동하는 경로를 시뮬레이션한 선행연구의 모형을 기본으로 하여 중국-중앙아시아 접경 지대 지형 위에 목초지 품질과 분포를 설정하고 정주 밀도와 인구, 그리고 이동 거리에 들어가는 비용을 변수로 입력할 수 있는 시뮬레이션 모형을 구축한다. 그런 다음 겨울에 목축 사회가 정주할 수 있는 저지대 지역에 무작위로 5,000개의 가상 사회를 점으로 배치하고, 겨울이 끝나면 목초지를 따라 이동하고 그 경로를 누적시키는 작업을 500번 수행한다. 시뮬레이션 1회 실행이 한 계절을 의미한다면 오차 범위를 감안해도 몇 세기에 걸친 시간을 가상으로 돌려보게 되는 것이다. 이 누적 경로를 종합하면 실크로드 회랑 지역의 고지대에 무수히 많은 돌기와 선이 점차 이어지면서 자연스럽게 특정한 경로와 중심지가 나타난다. 논문은 실험 대상의 지역에 실제 발견된 618개의 유적지 중 258개가 모형이 생성한 지점과 겹치는 것을 보이며 모형의 타당성을 입증하고 있다. 이렇게 일치한 258개 유적지가 임의로 생성되기 어렵다는 사실을 통계분석으로 입증하여 모형의 신뢰도를 한층 높이고 있다. 또한 논문에서는 시뮬레이션으로 500번 수행하는 중간 과정을 단계별로 보이며 고지대 경로가 생성되는 모습을 지도에 담아 이해를 돕고 있는데(〈그림 10〉 참고), 이러한 목축성 이동 경로 패턴은 오랜 시간 흐름에 따라 의도하지 않고도 자연스럽게 생성될 수 있는 결과임을 보이고 있다. 이 시뮬레이션 기반 연구는 오래 전부터 확립되어 온 중앙아시아 산악 유목민의 이동성 패턴에 따라 고지대 실크로드 네트워크가 서서히 출현했을 것이라 주장한다.



〈그림 10〉 톈산산맥과 타림분지 고지대에서 목축성 이동누적경로 시뮬레이션을 수행한 결과를 단계별로 표현한 지도. (a)는 1번 수행, (b)는 10번 수행, (c)는 500번 수행한 결과. 점과 선의 색깔이 진할수록 누적된 이동 횟수가 많은 것으로 시뮬레이션 수행 회수가 많아질수록 선과 점을 잇는 경로가 선명해짐을 알 수 있다. 기본적인 목축 이동양식과 규칙만 설정하여도, 이러한 실크로드 고지대 연결 네트워크가 시간에 따라 자연스럽게 형성될 수 있음을 보이고 있다.(출처 : 논문 A8의 Figure 2.)

34) Frachetti, Michael D., C. · Smith, Evan · Traub, Cynthia M. · Williams, Tim, "Nomadic ecology shaped the highland geography of Asia's Silk Roads.", *Nature* 543: 7644, 2017, p. 193.

이 연구의 기술적 면모를 좀 더 살펴보면 다른 시사점도 찾을 수 있다. 우선 잘 알려진 GIS 소프트웨어인 'ArcGIS'[35]를 이용했다는 점인데, 실제 지형을 그릴 뿐만 아니라 모형에서 흐름을 시뮬레이션하기 위한 핵심 부분을 위 소프트웨어가 제공하는 알고리즘을 이용했음을 밝히고 있다. 즉 이 시뮬레이션은 실제 물리적 흐름과 이동을 표현하는 엔진에 해당하는 복잡한 부분은 직접 만들지 않고 기존의 만들어진 모듈을 가져와 사용하여 구현의 난이도를 크게 낮췄다. 이처럼 일견 복잡해 보이는 시뮬레이션도 연구자들이 개념만 명확히 정의하고 설계할 수 있다면, 비교적 쉽게 만들어 볼 수 있다는 점을 시사하고 있다. 이와 비슷하게 행위자기반모형(ABM)에서 많이 사용하는 넷로고(NetLogo)[36]라는 소프트웨어가 있지만, 실제 지형을 그대로 활용하기에는 한계가 있는 구조이다. 이 연구에서 사용한 것처럼 실제 지형과 당시 시대 환경을 쉽게 설정하여 시뮬레이션 할 수 있는 도구를 사용한다면 더욱 수준 높은 연구가 많이 나올 수 있을 것으로 보인다.

이외에도 역사를 바라보는 관점에 대한 새로운 시도가 대두되고 있다. 복잡계과학이라는 통계물리학의 사회과학적 변용은 40년이 되는 시간에 걸쳐 서서히 인문사회 영역에 골고루 영향력을 보이고 있다. 에릭 클라인(Eric Cline)의 유명한 저서, *1177 B.C.: The Year Civilization Collapsed*(한국어판 제목 『고대 지중해 세계사』)는 고고학과 역사학에서도 본격적으로 그러한 해석을 도입한 대표적 사례이다. 클라인은 고대 지중해 세계의 청동기 문화가 왜 갑작스레 무너졌는지에 대해 단일한 인과관계를 설정하지 않는다. 오히려 그 어떠한 원인도 결정적이지 않으며, 복합적이고 서로 얽혀 영향을 끼쳤을 것이라는 다소 애매하게 보일 수 있어 비판의 소지가 있는 결론을 조심스럽게 제시하고 있다.[37] 그러나 복잡계적인 관점을 역사학의 큰 담론 중 하나에 적용했다는 점에서는 새로운 전환점이 될 여지가 보인다. 다만 비판의 목소리도 일리가 있는 바, 이러한 결론을 시뮬레이션을 통해 그 가능성을 설명할 수 있다면 더욱 훌륭한 해석으로 보강될 수 있을 것이다. 이처럼 시뮬레이션 방법론에서 복잡계 시스템은 기본적으로 고려해야 할 사항이다. 오늘날과 과거의 인간 사회 모두를 복잡계라고 본다면, 이를 기반으로 한 시뮬레이션은 인류학, 고고학, 그리고 역사학이 그동안 모은 수많은 퍼즐 조각을 모아 맞춰볼 수 있는 방법을 제공해 줄 수 있을 것이다.

이렇게 데이터의 구축에서 시작하여, 분석을 거치고 모형을 만들어 시뮬레이션하는 부분까지 이어지는 학계의 최근 흐름을 논문을 통해 살펴보았다. 이 과정에서 데이터과학의 방법론이 곳곳에 적용되어 있음을 확인할 수 있었다. 대체로 고고학에 치중된 연구 결과는 아쉬운 면이 있지만 다른 한편으로는 최고 수준의 자연과학 학술지에 수용될 수 있는 방법론은 인문사회학에도 크게 활용될 여지가 있음을 보여주기도 한다. 특히 고고학과 밀접한 관계인 역사학은 비교적 쉽게 데이터과학의 방법론을 적용하고 성과를 올릴 가능성이 매우 크다. 물론 이를 위해서는 문헌 연구 중심의 역사학에서 더욱 유용하게 활용될 수 있는 텍스트 분석 기술이 필수적이라 할 수 있겠다. 본 글에서는 다루는 논문들은 모두 고고학을 기반으로 하고 있어 텍스트 분석

35) Esri 사의 지리정보시스템(GIS) 소프트웨어. 지도를 포함해 지리 데이터 전반을 광범위하게 다룰 수 있는 소프트웨어이자 플랫폼으로 다양한 제품군이 있다.

36) NetLogo는 사회 현상을 ABM 형태로 시뮬레이션할 수 있는 프로그래밍 언어로 1999년 노스웨스턴 대학에서 만들어 지금까지 널리 쓰이고 있다. 오픈 소스로 공개되어 있어 누구든 자유롭게 이용할 수 있다. https://ccl.northwestern.edu/netlogo/ 참고(검색일: 2020.02.12.)

37) 에릭 클라인(저), 류형식(역), 『고대 지중해 세계사』, 소와당, 2017, 284~291쪽.

사례를 다루지 못한 아쉬움이 있다. 그러나 문헌 사료 연구 또한 결국 데이터과학을 이용하려면 데이터를 數로 변환해야 하고, 그 숫자에 의미를 담아 분석하게 되므로 지금까지 설명한 방법론을 모두 거치게 되는 것은 사실이다. 차후 연구에서 텍스트 분석에 집중하여 역사 연구에 효과적인 데이터과학 방법론을 살펴보도록 할 예정이다.

마지막으로 살펴볼 사례는 데이터과학을 기본 도구로 삼아 학문간 정보 교류의 수단으로 이용하고, 이를 통해 많은 연구자와 다양한 학문이 얽힌 거대한 프로젝트이다. 앞으로의 역사학, 그리고 고대 문명 연구가 어떻게 초학제간 연구로 발전해 나갈 것인지를 다음 장에서 살펴보도록 하겠다.

V. 인류의 근원적 호기심을 찾아가는 거대 학문의 흐름: 학제간 협력 연구와 글로벌 프로젝트

이제 맨 처음 소개했던 두 건의 대규모 공동연구를 주목할 때가 되었다. 유럽인의 기원을 다룬 두 연구(A9・A10)는 여러 가지 공통점을 보이는데, 모든 고대사 연구자들에게 관심을 끌 만한 문명의 이주, 전파, 정착에 관한 주제를 다루고 있으며 양대 과학 저널인 『네이처』와 『사이언스』에 각각 실렸다는 유사점이 있다. 그중에서 가장 큰 공통점은 데이비드 앤서니와 그의 초원 가설로, 그는 4년의 시차를 둔 두 연구에 모두 연구자로 참여하였다. 이처럼 고고학・역사학에서 뜨거운 이슈로 논쟁 중인 가설을 두 개의 서로 다른 연구로 입증했다는 점은 주목할만하다. 그러나 여기서는 연구 내용인 초원 가설 자체의 타당성보다는 연구의 규모와 협력이 어떻게 진행되었는지에 초점을 맞추고자 한다. 특히 후속 연구가 몇 배 이상 커지고 복잡해졌지만 더 높은 엄밀함을 추구하고 이를 통합하여 완결된 설명을 제시하고 있는 점은 인상적이다. 역사학의 글로벌 프로젝트가 어떻게 가능한지를 이 두 논문(A9・A10)으로 살펴보도록 한다.

(A9) 「초원지대에서 시작된 대규모 이주가 유럽의 인도유럽어의 근원이었다.」[38]

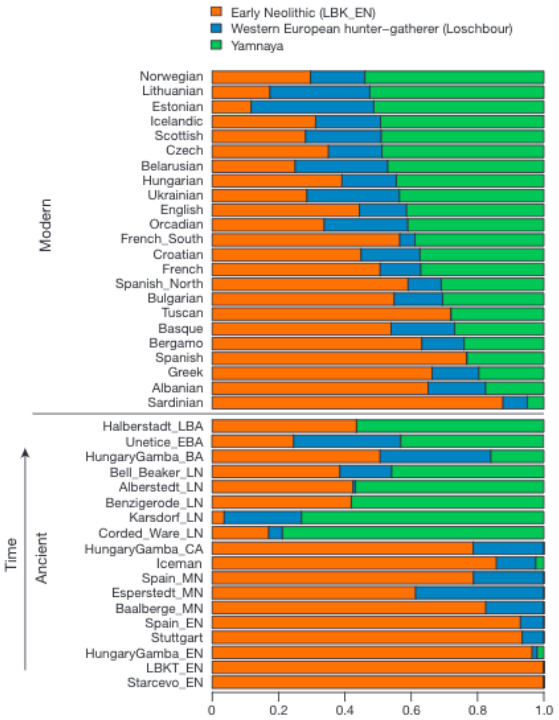
이 연구는 유럽 서부에서 러시아에 이르기까지 각지에서 발견된 94명의 고대 유럽인의 인골에서 DNA를 추출해 그 기원을 분석한 전형적인 의학・분자생물학 논문이다. 논문의 대부분은 유전자 분석 방법론에 대한 설명과 통계 분석결과로 구성되어 있으며, 역사적 의미와 고고학적 논의는 마지막 부분에 집중되어 있다. 연구의 주된 논지는 초원지대와 농경문화를 대표하는 유럽의 각 지역에서 발굴된 古인골의 유전자를 추출하고 이를 오늘날 유럽인의 유전자와 비교하여 인도유럽어족 인구 이동이 어느 방향에서 진행되었는지를 보이려는 데 있다. 인도유럽어족의 기원에 대해서는 오래전부터 두 가설이 대립하고 있는데, 약 8,500년 전 아나톨

38) Haak, Wolfgang, et al., *op. cit.*, 2015, p. 207.

리아 반도에서 유럽으로 넘어와 신석기 농경문화를 형성한 사람들이 인도유럽어족의 기원이라는 '아나톨리아 가설'과 러시아 스텝 지역의 유목민의 이주를 기원으로 보는 '초원 가설'이 그것이다.[39]

이 논문은 후기 신석기시대에 초원 지역에서 대규모 이주가 발생했고, 고대 중부 유럽인의 상당수 - 최대 75%까지 - 가 초원지대에서 온 이주민으로 교체되었을 것이라는 결론을 DNA 분석을 통해 내리고 있다. 이는 8,500년 전 아나톨리아 반도에서 유럽으로 넘어와 신석기 농경문화를 형성한 사람들이 인도유럽어족의 기원이라는 '아나톨리아 가설'을 반박하는 결과로, 논문은 약 4,500년 전 초원지대의 핵심 지역으로 추정되는 러시아의 얌나야(Yamnaya)로부터 대규모 이주가 있었을 가능성이 높음을 보이며, 이를 뒷받침하는 초원 가설의 타당성에 대해 설명하고 있다. 고대 인골에서뿐만 아니라 현대 유럽인의 DNA에서도 초원지대의 기원을 볼 수 있다는 결과는 〈그림 11〉로 확인할 수 있다.

오늘날 고인골의 유전자 분석을 통한 연구는 빈번하기에 그 방식만으로는 크게 놀랄 수준이 아닐 지도 모른다. 그러나 이 연구는 연구 자료와 연구자 그룹의 거대한 규모와 함께, 서로 다른 분야에서 다양한 방식으로 검증하고 이를 체계적으로 조합하여 신빙성 높은 결론을 얻었다는 점에서 주목할 만하다. 수많은 다방면의 전문가가 복잡한 협력 작업으로 인류의 달 착륙이라는 거대한 프로젝트를 성공시킨 것처럼, 역사 연구 또한 이처럼 큰 규모의 학제간 협업이 가능하다는 것을 이 연구를 통해 확인할 수 있다.

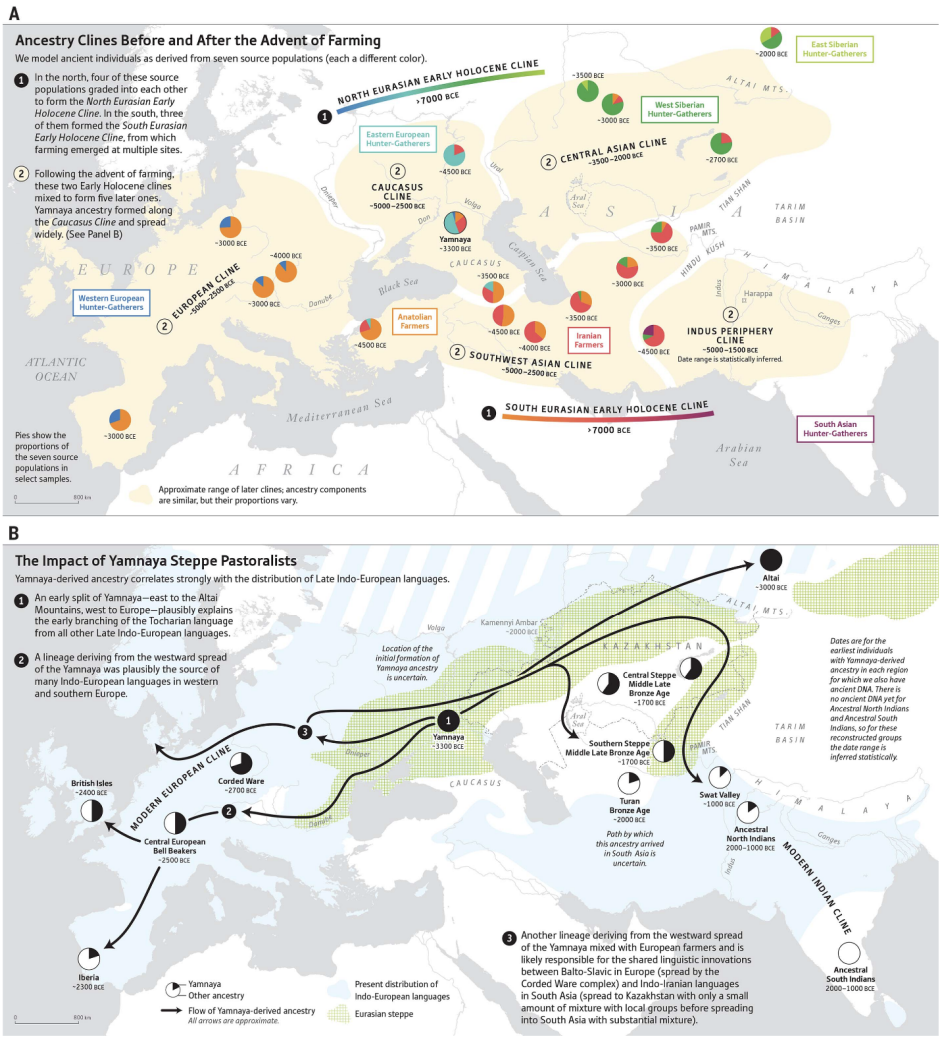


〈그림 11〉 지표 유전자 분포를 초기 신석기인, 서유럽 수렵-채집인, 얌나야(초원)인으로 구성으로 구분한 표.
녹색은 초원지대인 얌나야 지역 고대 인골의 유전자를 의미한다. 가로선 아래쪽 그룹은 고대인의 유전자 분포로, 초기에는 주황색 일색의 신석기 유럽인의 유전자만 보이다, 갑자기 녹색의 초원지대 유전자 그룹이 크게 등장하는 것을 볼 수 있다. 즉 특정 시점에 급격한 인구 변동이 있었음을 알 수 있다. 가로선 위쪽 그룹은 오늘날 유럽인의 유전자 분포로 얌나야 지역의 유전자가 오늘날 유럽인에게는 대부분 보인다.
(출처 : 논문 A9의 Figure 3.)

39) Pereltsvaig, Asya · Lewis, Martin, *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics*, Cambridge University Press, 2015, pp. 39~52.

(A10) 「남부, 중부 아시아의 인구집단 형성」[40]

2019년 『사이언스』에 실린 이 논문은 그 규모가 더욱 거대해졌는데, 이전 연구(A9)보다 더욱 광범위한 지역의 인골을 대상으로 한 연구이다. 이번에는 523명의 중앙아시아와 남아시아 지역 고대인의 DNA를 이용하는데, 유럽과 러시아 이외에도 중앙아시아와 인도까지 포함하여 당시 이주의 양상이 어떠하였는지 더 큰



〈그림 12〉 (A) 신석기시대 유럽의 농경 시작 시점의 인종적 분포. 초록색 계열과 붉은색 계열로 나뉜 집단이 유럽 전역에 있었음을 알 수 있다. (B) 초원 지대에서 유럽과 이란, 북인도 지역까지 확산된 이주 양상. 현대의 인도유럽어족 인구 분포(하늘색)과 유사함을 알 수 있다.(출처 : 논문 10의 Figure 3.)

40) Narasimhan, Vagheesh M., et al., *op. cit.*, 2019.

시각으로 바라보고자 하는 목적으로 진행되었다. 이 연구에서도 DNA 분석은 핵심 도구로 쓰이고 있으며, 이를 이용해 고고학 및 인류학 측면에서 연구되어 온 이주 양상이 실제 유전자 증거와 어떻게 일치하는지를 살펴보고 있으며, 그 규모를 보다 확대한 것이 앞선 연구(A9)와의 차이점이다. 117명의 공동저자 중 많은 수가 유라시아 각 지역의 고고학과 인류학 전문가들이라는 점은 연구의 국제성을 잘 드러내고 있다. 이 연구는 거대한 규모답게 300페이지가 넘는 연구 보강자료를 포함하고 있으며, 보강자료에서는 각 연구자가 맡은 지역에서 고고학적 증거와 유전자 분석 결과가 일치하는 경향이 있는지에 대해 상세히 다루고 있다.

연구의 핵심은 다음과 같다. 신석기 후반, 아시아 중부 지역은 수렵, 채집 위주의 인구 집단이 거주했고, 아시아 남부에서는 농경 집단이 존재했는데, 둘은 서로 유전적으로도 구분되어 있었다(〈그림 12〉 A 참고). 아시아 남부에서 유럽으로 농경이 확산되었을 것으로 보이며, 이 시기의 아나톨리아 반도를 포함한 서남아시아 지역에 거주한 사람들과 유럽인의 유전자는 상당수 일치함을 알 수 있다. 이러한 양상은 기원전 3천년기에 급격히 변화를 맞이한다. 흑해와 카스피해 사이에 위치한 오늘날 러시아 암나야 지역의 초원지대에 살았던 사람들이 유럽과 중앙아시아로 빠르게 이주한 사실을 유전자 분석을 통해 확인할 수 있다. 이 초원지대의 사람들은 과거 중앙아시아에 살던 사람과도 달랐으며, 고대 유럽에 살던 농경민과도 유전적으로 큰 차이가 있었다. 〈그림 12〉에서 이러한 확산 이전과 이후의 변화가 어떠했는지 알 수 있다. 암나야 초원 유목민의 확산으로 유럽인의 상당수는 이들의 유전자를 큰 비율로 가지게 되었으며, 비슷한 비율을 당시 카자흐스탄과 파미르고원, 투란에 이르는 중부 초원지대에 살았던 사람들도 지니게 되었음을 확인할 수 있다. 이 유전자는 인도 북서부까지는 나타나나, 인도 남부 지역에서는 전혀 보이지 않아 인도유럽어족의 분포와 유전적 인구 분포가 상당히 유사하다는 것 또한 알 수 있다. 즉 이 연구는 고인골의 유전자 분포 현황과 변화 시기를 초원 가설로 설득력 있게 설명할 수 있다는 것을 보이면서, 더 나아가 인도유럽어족을 포함한 고대 유럽인의 이주가 어떻게 일어났는지도 포괄적으로 제시하고 있다. 이 논문은 이미 거대한 프로젝트를 더욱 크게 만들어 그 답을 찾아가는 과정을 볼 수 있는 좋은 사례로 볼 수 있다.

Ⅵ. 맺음말

지금까지 살펴본 논문들은 고고학과 역사학 주제로 과학 학술지에 실린 것으로, 모든 연구가 데이터과학의 방법론을 활용하고 있다. 이는 데이터과학은 특별한 기술이 아닌, 자연과학에서 일반적인 통계적 분석 방법의 확장으로 사용되고 있음을 의미한다.

그렇다면 역사학자는 어떻게 데이터과학을 활용할 수 있을까? 자연과학과 일부 사회과학에서 이미 자연스럽게 사용한다 해도, 일견 복잡해 보이는 수식과 방법론을 연구에 활용하기란 쉽지 않은 것이 사실이다. 그리고 방법론을 이해하더라도 적용은 또 다른 문제가 된다. 통계와 데이터처리에 익숙하지 않은 상황에서 굳이 시간을 들여 확실하지 않은 방향으로 연구를 진행할 필요가 있는가? 이는 자연스럽게 귀결되는 질문이다.

다행히도 활용 방안에 대한 해결책은 이미 가까이 있고, 이미 살펴본 바 있다. 바로 협력 연구가 그것이다. 과학계에서 일반화된 협력 연구를 역사학에서도 주도적으로 진행해볼 필요가 있다. 새로운 도구의 강력함 때문에 그 사용법이 낯선 역사학자들은 데이터과학이라는 도구를 손에 쥔 匠人을 경계할 수 있다. 그러나 약자는 도구를 들고 있는 사람들이다. 그들은 데이터가 없으면 아무것도 할 수 없으며, 기술적인 분석만으로는 제대로 된 연구를 할 수 없다는 것을 잘 알고 있는 사람들이다. 역사학계에서 사료의 데이터화 과정부터 모든 통계 분석의 해석에 관여하지 않는다면 이 분야에 발길을 돌릴 수밖에 없는 사람들이다. 아무리 뛰어난 데이터과학자가 있더라도 의료 데이터를 의사의 도움 없이 해석할 수 없고, 치료에 쓸 수 없는 것과 같은 이치이다. 손길은 이 분야의 전문가인 역사학계에서 내밀어야 한다.

이러한 연구 형태를 강조하고자 새로운 이름을 붙이고자 한다. '데이터역사과학(Data History Science)'이라는 명칭은 조금 어색하고 단어 배열순서도 혼동의 여지가 있으나, 그 특징을 드러낼 수 있는 가장 단순하면서 효과적인 이름이다. 데이터과학은 이 분야에서 분석 도구이자 동시에 소통의 도구로 활용될 수 있다. 전통적인 인문학자, 특히 문헌에 익숙한 역사학자들은 이 소통의 시작으로 '데이터 리터러시(Data Literacy)'로 불리는 데이터 문해력을 체득해 나갈 필요가 있다. 본인이 잘 알고 있는 사료를 숫자로 바꿔 이해하는 것부터 출발하고, 그것을 표와 그래프로 표현하는 것에 익숙해지는 것이 출발점이다. 다행히도 데이터과학자들은 본인들이 모르는 데이터를 다루는 데 익숙하기에, 데이터를 잘 아는 전문가와 협력하는데 이골이 난 사람들이다. 협력연구를 하기에 이처럼 좋은 파트너 집단도 드물 것이다.

특히 고대 문명을 연구하는 데 있어 데이터역사과학은 큰 도움이 될 수 있을 것이다. 이 글에서 다룬 10개 논문 중 8개가 고대사와 관련 있는 것이었기에 이미 그러한 가능성은 자연스럽게 체감하였을 것으로 생각한다. 그 중에서도 마지막에 다룬 논문(A10)은 고대사에서 서로 다른 지역과 시대, 문화 연구를 하나의 주제와 데이터 분석을 통해 연결할 수 있고, 각각의 문화 연구로도 수준 높은 결론을 낼 수 있다는 점에서 주목할 만하다. 이미 고고학을 통해 고대 문명의 실마리는 학제간 연구로 이어지고 있다. 역사학도 이러한 연구가 가능하고, 또한 충분히 매력적인 자료와 연구 성과 또한 준비되어 있다. 데이터역사과학을 협력과 학제간 연구의 출발점으로 삼아 과거를 새로운 관점으로 해석해볼 수 있기를 바란다. 새로운 연구의 키는 역사학계가 쥐고 있다.

〈참고문헌〉

데이비드 W. 앤서니(저), 공원국(역), 『말, 바퀴, 언어: 유라시아 초원의 청동기 기마인은 어떻게 근대 세계를 형성했나』, 에코리브르, 2015.

에릭 클라인(저), 류형식(역), 『고대 지중해 세계사』, 소와당, 2017.

이혜림, 「국가 고고학 데이터 디지털 아카이브 개발을 위한 연구」, 『한국기록관리학회지』 18: 2, 2018.

Anthony, David W., *The horse, the wheel, and language: how Bronze-Age riders from the*

Eurasian steppes shaped the modern world, Princeton University Press, 2010.

Candela, Leonardo · Castelli, Donatella · Manghi, Paolo · Tani, Alice, "Data journals: A survey.", *Journal of the Association for Information Science and Technology* 66: 9, 2015.

Chandler, Tertius, *Four thousand years of urban growth: an historical census*, The Edwin Mellen Press, 1987.

Cline, Eric, *1177 B.C.: The Year Civilization Collapsed*, Princeton University Press, 2014.

Corrales, David Camilo · Ledezma, Agapito · Corrales, Juan Carlos, "A conceptual framework for data quality in knowledge discovery tasks (FDQ-KDT): A Proposal.", *Journal of Computers* 10: 6, 2015.

Frachetti, Michael D., C. · Smith, Evan · Traub, Cynthia M. · Williams, Tim, "Nomadic ecology shaped the highland geography of Asia's Silk Roads.", *Nature* 543: 7644, 2017.

Guedes, Jade d'Alpoim · Bocinsky, R. Kyle, "Climate change stimulated agricultural innovation and exchange across Asia.", *Science Advances* 4: 10, 2018.

Haak, Wolfgang · Lazaridis, Iosif · Patterson, Nick · Rohland, Nadin · Mallick, Swapan · Llamas, Bastien · Brandt, Guido, et al., "Massive migration from the steppe was a source for Indo-European languages in Europe.", *Nature* 522: 7555, 2015.

Hosner, Dominic · Wagner, Mayke · Tarasov, Pavel E · Chen, Xiaocheng · Leipe, Christian, "Spatiotemporal distribution patterns of archaeological sites in China during the Neolithic and Bronze Age: An overview.", *The Holocene* 26: 10, 2016.

Kohler, Timothy A. · Buckland, Philip I. · Kintigh, K. W. · Bocinsky, R. K. · Brin, A. · Gillreath-Brown, A. · Ludäscher, B. · McPhillips, T. M. · Opitz, R. · Terstriep. J., "Paleodata for and from archaeology.", *PAGES Magazine* 26: 2, 2018.

Leipe, C. · Long, T. · Sergusheva, E. A. · Wagner, M. · Tarasov, P. E., "Discontinuous spread of millet agriculture in eastern Asia and prehistoric population dynamics.", *Science Advances* 5: 9, 2019.

Modelski, George, *World cities: -3000 to 2000*, Faros, 2000.

Narasimhan, Vagheesh M. · Patterson, Nick · Moorjani, Priya · Rohland, Nadin · Bernardos, Rebecca · Mallick, Swapan · Lazaridis, Iosif, et al., "The formation of human populations in South and Central Asia.", *Science* 365: 6457, 2019.

Ortman, Scott G. · Cabaniss, Andrew HF · Sturm, Jennie O. · Bettencourt, Luís MA, "Settlement scaling and increasing returns in an ancient society.", *Science Advances* 1: 1, 2015.

Penny, Dan · Zachreson, Cameron · Fletcher, Roland · Lau, David · Lizier, Joseph T. · Fischer, Nicholas · Evans, Damian · Pottier, Christophe · Prokopenko, Mikhail, "The demise of Angkor:

Systemic vulnerability of urban infrastructure to climatic variations.", *Science advances* 4: 10, 2018.

Pereltsvaig, Asya · Lewis, Martin, *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics*, Cambridge University Press, 2015.

Reba, Meredith · Reitsma, Femke · Seto, Karen C., "Spatializing 6,000 years of global urbanization from 3700 BC to AD 2000.", *Scientific Data* 3, 2016.

Romanowska, Iza · Stefani Crabtree · Benjamin Davies · Kathryn Harris, "Agent-based Modeling for Archaeologists. A step-by-step guide for using agent-based modeling in archaeological research (Part I of III)." *Advances in Archaeological Practice* 7, 2019.

Turchin, Peter · Currie, Thomas E. · Whitehouse, Harvey · François, Pieter · Feeney, Kevin · Mullins, Daniel · Hoyer, Daniel, et al., "Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization.", *Proceedings of the National Academy of Sciences* 115: 2, 2018.

Turchin, Peter · Whitehouse, Harvey · François, Pieter · Hoyer, Daniel · Alves, Abel · Baines, John · Baker, David, et al., "An Introduction to Seshat: Global History Databank.", *Journal of Cognitive Historiography*, 2019.

Whitehouse, Harvey · François, Pieter · Savage, Patrick E. · Currie, Thomas E. · Feeney, Kevin C. · Cioni, Enrico · Purcell, Rosalind, et al., "Complex societies precede moralizing gods throughout world history.", *Nature* 568, 2019.

〈d'Alpoim Guedes, Bocinsk의 연구논문 소스코드 사이트〉 (Github)
https://github.com/bocinsky/guedesbocinsky2018

〈CoMSES network〉 웹사이트 https://www.comses.net/

〈NetLogo〉 웹사이트 https://ccl.northwestern.edu/netlogo/

〈세스헤트(Seshat) 〉 웹사이트 http://seshatdatabank.info/

〈Scientific Data〉 웹사이트의 저널 소개 (한국어) https://www.natureasia.com/ko-kr/scientificdata/

* 이 논문은 2020년 2월 21일에 투고되어,
2020년 4월 3일까지 심사위원이 심사하고,
2020년 4월 7일에 편집위원회에서 게재가 결정되었음.

▌Abstract ▐

Data History Science: The Gateway to the Big Questions of the Ancient Worlds

Ghim, Gwanglim*

In today's scientific world, history subjects have been studied in various methods over history, and also meaningful researches have been published. This paper shows that reliable historical data and rigorous data analysis are fundamental to the background of such research. Data Science makes critical points to the studies. Various articles about historical and archaeological topics published in the science journals are reviewed to examine how data science can be powerful utilities in history. This paper explains the 'data-based simulations' methodology that allows them to observe changes over time by changing situations in the past, and it can be useful in the studies of ancient civilizations. In large-scale global research projects, where interdisciplinary coworking is essential, data science is a key feature not only as a method but also as communication. For this purpose, the term "Data History Science," a blending of history and data science, is newly proposed.

[Key Words] History, Data Science, Data History Science, Databases, Data Analysis, Archaeology, Complex systems, Simulations, Ancient Studies, Interdisciplinary Studies, Digital Humanities

* Ph.D. process, Dankook University