EMB00003ae06aad

졸업작품/논문 최종보고서

2017학년도 제 2학기

제목 : 기온 분석을 통한 조류 인플루엔자 바이러스 감염 발생 예측 모델

두건(2013310089)

2017년 11월 일

지도교수: 서명

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 계획(10) | 주제(20) | 개념(20) | 상세(30) | 보고서(20) | 총점(100) |
|  |  |  |  |  |  |

\* 지도교수가 평가결과 기재

* **요약**

2016년 11월부터 시작된 조류 인플루엔자로 인해 대한민국은 산란계의 떼죽음으로 전례 없는 계란 값 폭등을 겪었다. 대부분 음식에 자주 사용되는 계란 원가가 폭등하자 외식업계는 음식 가격을 인상하거나 계란 사용을 일시 중지하는 등 전국적인 큰 혼란이 빚어졌다. 본 논문에서는 데이터 분석 기법을 이용해 조류 인플루엔자 바이러스와 해당 지역의 기온의 상관 관계를 밝혀내고자 한다. 기온 외에도 습도, 강수량 등 외부 환경 요인을 늘릴 수 있으나 데이터 수집의 한계로 본 논문에서는 기온과의 관계만 파해지고자 한다. 구체적으로는 회귀분석 기법을 이용해 조류 독감과 외부 기온의 상관관계를 파악한 후 조류 인플루엔자 발생과의 연관성을 측정하고자 한다. 조류 인플루엔자 발생 데이터는 농림축산검역본부 사이트에서 공개 데이터로 얻을 수 있고 기온 데이터는 기상정보개방포털 사이트에서 데이터를 얻을 수 있다. 본 논문을 통해 (1) 조류 인플루엔자 바이러스와 기온의 상관관계 측정, (2) 측정을 통해 향후 연구 혹은 활용방안으로써 활용할 수 있을 것이다.

* **서론 (6페이지 이내)**

1. **제안 배경 및 필요성**

2016년 11월부터 발생하기 시작한 조류 인플루엔자는 대한민국 사회에 매우 큰 영향을 끼쳤다. 이전부터 조류 인플루엔자로 인한 피해는 있었으나 그 여파가 일반 소비자 식탁에까지 미친 건 이번이 처음이다. 조류 인플루엔자로 인한 산란계 수 급감으로 계란의 공급이 현저히 감소했고 계란 값은 며칠만에 두 배 이상으로 치솟았다. 계란을 이용한 음식이 많은 한국인의 밥상 특성상 그 영향은 이루 말할 수 없을 정도로 거대했다. 2017년 3월 기준 살처분된 가금류의 수는 3000만 마리 이상으로 기존 사육되는 가금류 수의 20%를 차지한다(정빛나, 2017). 이렇듯 조류 인플루엔자로 인해 경제적, 사회적 손실이 매우 큰 상황이지만 문제는 여기서 끝나지 않았다는 점이다. 조류 독감으로 심각한 피해를 본 양계장 업주들은 이를 예방하기 위해 인체에 해롭다고 판정된 살충제를 가금류에 투여했다. 결국 인간에게 해를 끼칠 수 있는 살충제 사용이 언론에 적발되면서 소비자는 계란 사용을 더욱 기피하게 되었다. 닭은 조류 인플루엔자로 인해 대량으로 살처분 되었고, 계란은 팔리지 않으니 양계업자들은 그야말로 사면초가에 빠진 것이다.

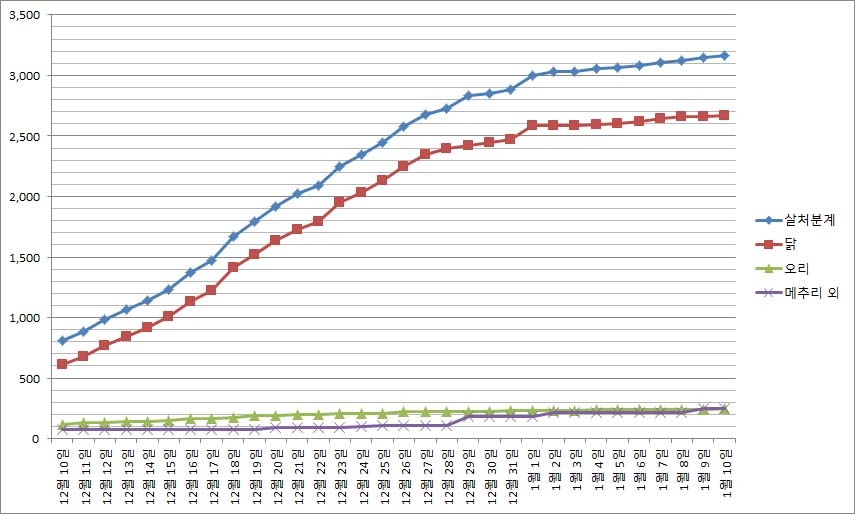


그림 2017년 1월10일까지의 살처분계 상황 그래프

이렇듯 조류 인플루엔자로 인한 일명 ‘계란 파동’은 11개월이 지난 지금에서야 서서히 안정을 찾기 시작했다. 하지만 살충제 파동으로 인한 계란에 대한 부정적인 시선은 여전히 존재하며 산란계가 급감한 양계업자들 또한 향후 산란계가 보충되기 전까지는 어려움이 지속될 것으로 보인다.

조류 인플루엔자로 인한 피해는 매년 발생하며 그 피해액도 상당하다. 실제 본 논문을 작성하기 위해 농림축산검역본부 사이트의 자료를 분석한 결과, 2013년부터 매년 조류 인플루엔자 발생 신고가 접수되었으며 그 횟수 또한 10건을 모두 넘어간다. 그림 2는 농림축산검역본부로부터 얻은 조류 인플루엔자 발생 신고 데이터를 분석하여 도출한 수치이다.

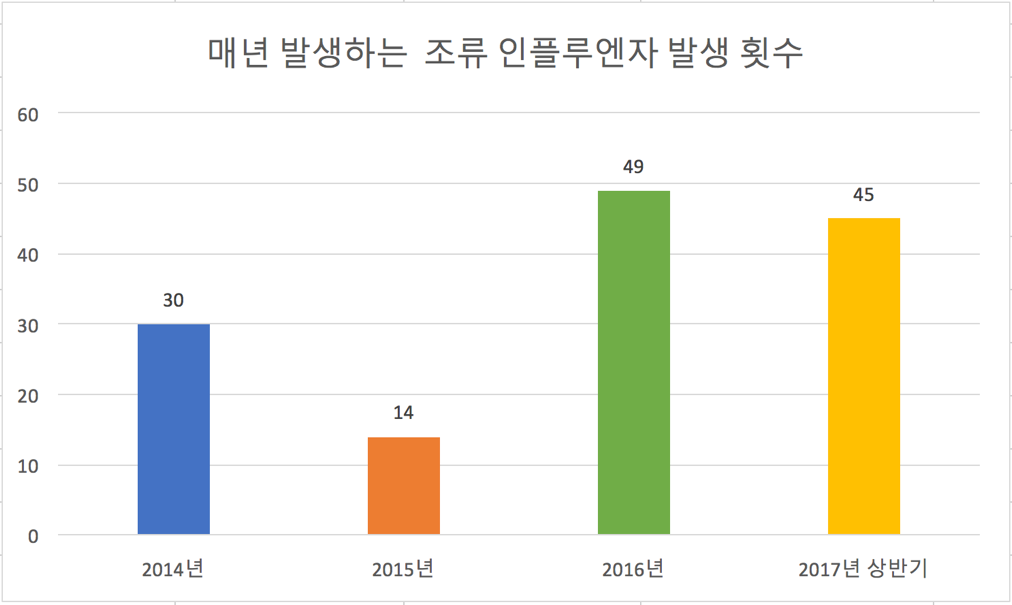


그림 2 2014년부터 2017년 상반기까지 발생한 조류 인플루엔자 횟수

사회적으로나 경제적으로나 한국 사회에 심각한 피해를 준 조류 독감을 예방하는 방법으로는 방역 작업이 전부일 정도로 기술적 지원이 안 되고 있는 상황이다. 4차 산업혁명 시대에 들어서면서 인간 전염병에 대한 예측 가법은 날이갈수록 발전하고 있지만 조류 인플루엔자와 관련된 예측 기법은 없다시피 하며 전파를 방지하는 방법도 이동 차량에 대한 검역을 강화하는 방법 밖에 없는 것이 현실이다. 시대는 발전하지만 조류 인플루엔자를 예측하거나 예방 시 구시대적인 방법을 사용하는 것이다. 본 논문을 작성하면서 조류 인플루엔자에 예방과 관련된 다른 논문을 검색했지만 방역 작업의 효율성을 높이는 방법 외의 내용은 찾아볼 수 없었다.

하지만 UN(이하 유엔) 산하의 LEAD Company(이하 리드 컴퍼니) 사례를 통해 조류 인플루엔자를 빅데이터 분석을 통해 예방할 수 있을 것이라 생각한다. 리드 컴퍼니의 주 목적은 통신사, 보험사 등에서 얻은 고객 빅데이터로 감염병의 전파 경로를 예측해 이를 사전에 예방하는 것이다. 4차 산업혁명 시대에 맞춰 인간 전염병을 대처하는 방법 또한 진화한 것이다. 이 방법을 조류 인플루엔자에 대입하여 그 전파를 방지할 수 있다면 한국 사회에 이번 조류 독감 및 살충제 계란 파동과 같은 손실을 예방할 수 있을 것이다. 본 논문이 주장하는 필요성은 바로 조류 인플루엔자의 예상 발생 지점을 예측하고 이를 방지에 사회적 손실을 줄이자는 것에 있다.

1. **논문의 목표**

본 논문의 최종 목표는 조류 독감을 1차 발생에 그치게 하는 것이다. 원천 방지가 가장 좋은 방법이지만 조류 인플루엔자는 철새에 의해 유입되므로 완전 예방은 불가능하다. 하지만 그 동안 조류 인플루엔자로 인한 피해가 커진 이유는 조류 인플루엔자가 다른 농장으로까지 전파되면서 피해가 기하급수적으로 늘어난 데에 있다. 그러므로 주변 환경 분석을 통해 조류 인플루엔자가 어떠한 환경의 지역으로 전파가 잘 되는지를 분석한다면, 특정 지역에 대한 검역을 강화함으로써 전파를 방지할 수 있고 이전과 같은 거대한 피해를 예방할 수 있다(김재만, 2015). 주변 황경으로는 습도, 날씨, 고도 등 다양한 데이터가 존해하지만 본 논문에서는 데이터 수집의 한계로 인해 기온만을 다루기로 하였다. 향후 더욱 다양한 데이터와 기법을 도입한다면 정확한 결론을 도출 할 수 있을 것이다.

1. **논문 전체 overview**

본 논문은 크게

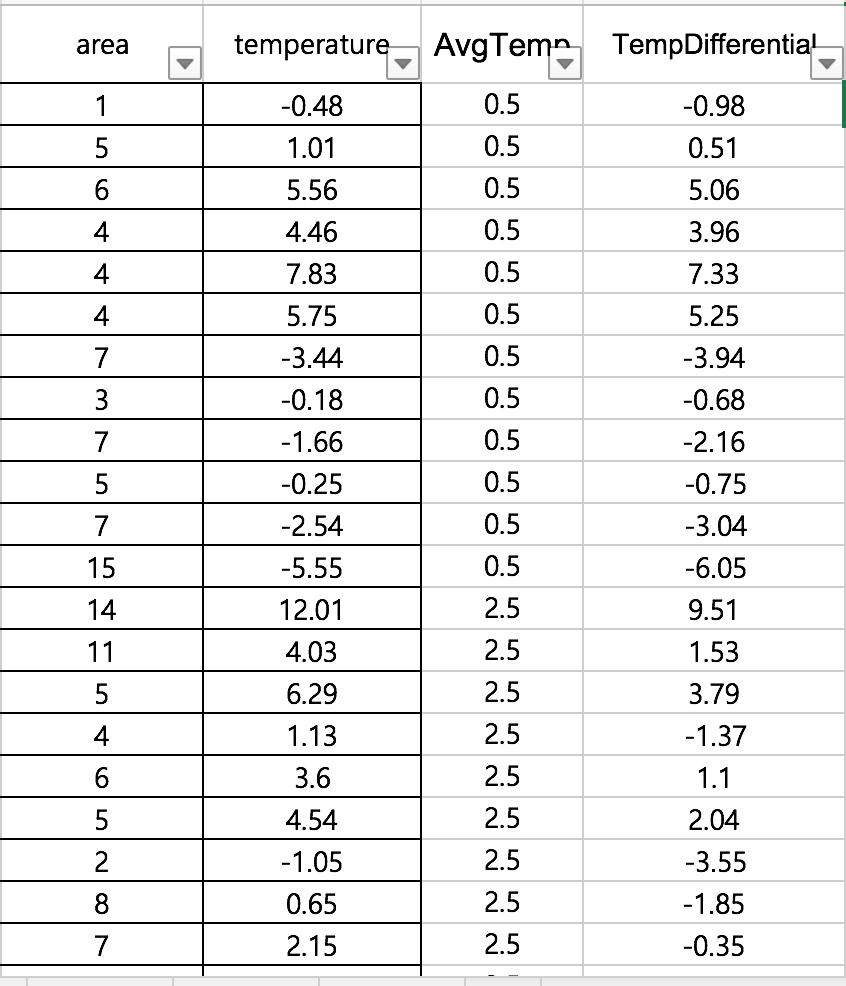
1. 데이터 수집, 생성 및 정리 상황
2. 정리된 데이터를 이용한 그래프 추출
3. 그래프와 수치 데이터를 이용해 회귀분석 기법을 실행하고, 기온과 조류 인플루엔자의 연관성 분석
4. 결론

이 4가지 부분으로 진행될 예정이다.

데이터 수집, 생성 및 정리 상황에서는, 조류 인플루엔자 신고 상황을 농림축산검역본부 사이트 (농림축산검역본부, 2017년)에서, 조류 인플루엔자 발생 지역의 기온 데이터 및 당시 전국 평균 기온 데이터는 기상자료개발포털 사이트 (기상자료개방포털, 2017년)에서 데이터를 얻을 수 있었다. 이들 사이트에는 상세한 데이터가 존재해 본 논문에서 필요로하는 데이터를 얻을 수 있었다. 그런 후 조류 인플루엔자 발생 지역과 날짜, 그리고 당시 해당 지역 기온 및 당시의 한국 평균 기온 데이터를 병합해 하나의 데이터 파일로 정리하였다.

데이터 파일의 열은 조류 인플루엔자 발생 구역(*area*), 당시 지역 기온(*temperature*), 당시 전국 평균 기온(*AvgTemp*) 그리고 기온과 전국 평균 기온의 차(*TempDifferential*)가 있다. 여기서 데이터 수집과 관련된 자세한 사항과 발생 구역(*area*)의 분류 기준은 제안 작품 소개에서 자세히 설명할 예정이다.

표 데이터 분석을 위해 작성된 데이터 파일 예시



데이터를 이용한 그래프 추출에서는, 통계학적 그래프 작성이 편리한 python 언어를 사용했다. Python에는 altair, pandas 등 여러 그래프 작성 라이브러리가 존재하는데 초기에는 altair를 이용해 그래프를 작성할 계획이었다. 하지만 어려움에 부딪히고 pandas를 이용하는 것이 더욱 용이하다는 것이 판단되어 이후 pandas 라이브러리를 이용해 그래프를 얻을 수 있었다 (PinkWink, 2015).

그렇게 그래프라는 시각적 데이터를 얻었고, 본 논문의 결론을 더욱 보강하기 위해 객관적 수치를 얻고자 하였다. 기온과 조류 인플루엔자 발생 지역 데이터 간의 연관성을 회귀분석 기법을 이용해 수치화하여 결론을 도출할 수 있었다. 처음 논문 의도와는 다르게 시각적 데이터로는 대략적인 연관성을 알 수 있었으나 자세한 연관성을 알아낼 수 없었다. 하지만 회귀분석 기법을 이용해 연관성을 분석하기 시작했고 multiple R이라는, 연관성을 객관적 수치로 나타내는 값을 얻을 수 있었다. 이 multiple R 값이 매우 높은 수치를 보여 기온과 조류 인플루엔자 간의 연관성이 높음을 알 수 있었다. Multiple R의 값과 관련된 글을 구현 및 결과분석에서 다룰 예정이다.

* **관련연구 (7페이지 이내)**
  1. 빅데이터와 전염병의 연관성

4차 산업혁명 시대가 도래하면서 종전에 존재하지 않았거나 미완이었던 이론이 빛을 보기 시작했다. 빅데이터도 그 중 한 예인데, 스마트폰의 보급과 모바일 인터넷으로 사람들은 언제 어디서나 데이터를 생성하기 시작했다. 폭발적으로 증가한 모바일 트래픽으로 인해 비정형적 데이터도 폭증하기 시작했고 이는 빅데이터 발전의 기폭제가 되었으며 여러 분야에서 빅데이터가 사용되기 시작했다.

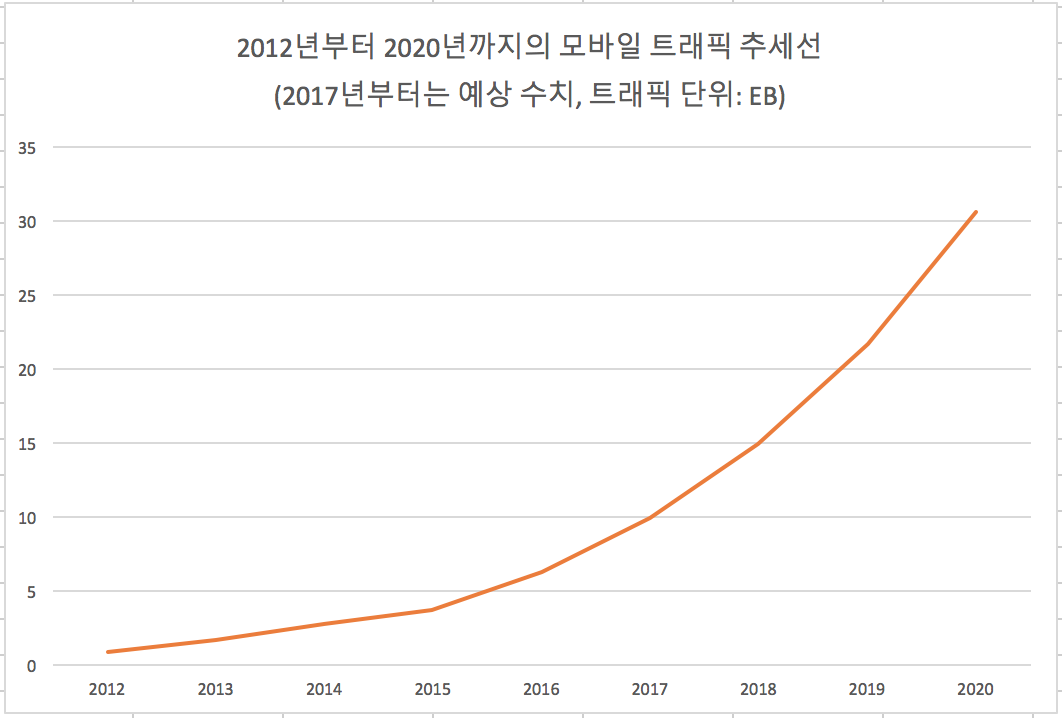


그림 2012년부터 2017년까지의 모바일 트래픽 추세선[[1]](#footnote-1)

이전까지 빅데이터를 이용한 연구 분야는 주로 정치, 마켓팅, 신제품 기획 등에 사용되어 왔다. 빅데이터를 이용하여 전염병을 다루는 연구는 매우 적게 이루어졌으며 이 또한 큰 주목을 받지 못한 채 활용이 못 되고 있다. 하지만 2015년 발생한 메르스 사태 이후 사람들은 기존 언론에 대해 불신하기 시작했고 SNS를 통해서 전염병과 관련된 정보를 얻기 시작했다. (허지윤, 2017) 이와 같이 국내의 경우 메르스 사태를 통해 빅데이터를 전염병과 연관 짓기 시작했고 관련 연구도 진행이 되어왔다. 하지만 SNS를 통해 괴담이 전파되면서 이에 대한 대중의 인식이 부정적으로 흘러갔고 SNS를 이용한 전염병 예방에 관한 연구도 힘을 얻지 못했다. 조류 인플루엔자의 경우에는 연구 자체가 매우 드물어 본 논문을 작성하기 위해 관련 연구를 검색하였지만 여의치 않았다. 실제로 RISS와 NDSL에서 “조류 독감 빅데이터” 키워드로 검색할 시 RISS에서는 관련 논문 하나, NDSL에서는 아예 관련 논문이 검색되지 않는다. 조류 인플루엔자와 비슷하게 가금류에게 발생하는 전염병으로는 구제역이 있다. 조류 인플루엔자와 빅데이터 간의 자료가 너무 적어 본 논문의 관련연구를 찾기 위해 “구제역 빅데이터” 키워드를 RISS에서 검색하였다. 그 결과 「빅데이터 기반 실시간 농축산 검역을 위한 확산경로 탐색과 역학조사 모델링 연구[1] 」 논문을 찾을 수 있었다. 이 논문에서는, 2010년 발생한 구제역과 농장 간 차량 이동을 연관 지어 특정 차량 방문 시 바이러스가 전염된다는 주제로 작성되어 있다. 이 논문에 의하면, 구제역 전파를 막기 위해 실시되는 기존 역학 조사는 특정 농장 방문 차량 리스트를 추출하거나 특정 차량의 방문 농장 리스트를 추출하는 기법을 사용하고 있다. 하지만 이 범례가 너무 커서 효율성이 떨어졌으며 실제로 조류 인플루엔자와 비슷한 빈도로 구제역이 발생해 농가에 피해를 주고있다. 이 논문에서는 농장을 방문한 특정 차량을 유형별, 목적별, 기간별 등으로 세세히 분류하고 특정 차량의 방문 농장을 유형별, 가축종류별, 기간별, 발병이력여부별 등으로 분류하여 분석하는, 보다 효율적인 시스템을 제시하였다. 이렇게 수집된 비정형적인 데이터를 빅데이터 기법을 분석하고, 이 중 유의미한 데이터를 추출하는 것에 방점을 두고 기존 시스템을 개선하고자 하는 이론을 제시하였다 (김종현(KimJong, 2014).

빅데이터 기법으로 구제역, 즉 전염병을 막는 것은 본 논문의 주제와 매우 흡사하다. 하지만 본 논문에서는 기온을 분석하여 전염병의 전파 지점을 예측하는 것에 방점을 두지만, 참고 논문에서는 농장 방문 차량에 대한 분류를 세분화시켜 분석 작업 효울성을 증가시키고 이를 통해 결론을 도출한다. 즉, 본 논문은 환경적 요인을 연구하지만 참고 논문은 인위적인 요인을 연구하는 것이다. 그러므로 본 논문과 참고 논문을 병행하여 사용한다면 구제역과 조류 인플루엔자 모두의 예상 발생 지역을 효과적으로 예측할 수 있을 것이라 생각한다.

* 1. 효과적인 논문 주제 전파 방식

한편, 또다른 논문인 「페이스북 이용을 통한 정보획득과 사회자본이 질병 예방행동의도에 미치는 영향[2]」에서는 메르스로 인해 페이스북을 이용한 전염병 관련 소식을 접하는 사람들이 늘고 있다는 것을 총 324명의 설문조사를 통해 도출하였다. 여기서 사람들은 빅데이터를 이용해 전염병과 관련된 소식을 들으면 오히려 전염병 예방에 소흘해 진다고 주장하고 있다. SNS 상에는 검증되는 않은 이야기와 괴담이 난무하면서 사람들은 SNS에서 생성되는 소식을 귀담아 듣지 않는 것이다. 그러므로 본 논문의 주장을 전파하기 위해서는 더욱 다양한 데이터와 분석 기법으로 객관적인 수치를 제시한 후, 여러 학회를 통해 본 논문의 아이디어를 전문가들에게 전하여 설득하고 실천하게 만드는 것이 효과적일 것이다.

* 1. 전염병과 빅데이터의 연관성에 대한 정부의 태도 변화 필요

「사회안전을 위한 빅데이터 활용의 재난대응 정책[3]」에 의하면, 현대사회는 고도의 전산화, 정보화가 진행되었는데도 불구하고 전염병과 관련해서 정부의 대응이 미흡하다고 지적하고 있다. (정영철최익수,, 2016) 빅데이터를 이용하여 바이러스의 패턴을 분석하고 이를 효과적으로 예방할 수 있음에도 불구하고 정부는 바이러스가 창궐한 이후에야 적극적인 대응을 실시하고 이마저도 검역 강화 등 구시대적인 방식을 이용하고 있다. 메르스 사태와 2016년 11월 발생한 조류 인플루엔자 사태는, 그 시작은 비록 우연이었을지 몰라도 정부의 미흡한 대응 방식이 한 몫한 것이다. 만약 정부가 빅데이터를 이용해 조류 인플루엔자 등의 전염병을 적극적으로 예방한다면 서론에서 서술한 것과 같은 재앙이 발생하지 않았을 것이다. 그러므로 정부는 더이상 바이러스가 발생한 이후에야 검역 강화를 하는, 구시대적인 방식에서 벗어나 빅데이터와 통계학의 다양한 분석 기법을 이용해 전염병을 예방해야 할 것이다.

* 1. 해외에서 입증된 빅데이터를 이용한 전염병 예방 효과

「Study on Big Data in Public Health, Telemedine and Healthcare[4]」에 의하면, 빅데이터를 이용한 공중보건 관리는 권장 사항이라 하고 있다. (Osterreich, 2016) 그러면서 각국의 정책, 그리고 국민들의 관심 등이 필요하다 주장하고 있다. 「Big Data Analytics for Healthcare[5]」에서는 빅데이터와 공중보건을 어떻게 하면 효과적으로 융합할 수 있는지를 서술하고 있다. 예컨대 빅데이터로 환자의 상태를 상세히 분석하고 관련 사례를 통해 어떠한 전염병에 감염될지를 예측하여 사람들이 병원비로 사용하는 액수를 낮출 수 있다는 것이다. (Jimeng Sun, 2013) 「Big Data for Health[6]」에서는 빅데이터를 이용하여 전염병에 발생 가능성을 예측하고 관련 백신 연구를 선행하여 효과적인 전염병 전파 저지를 하는 것을 서술하고 있다. 또한 사람의 DNA, 인체 구성 등을 빅데이터로 정리해 감염되기 쉬운 바이러스를 분류해서 전염병을 예방하는 것에도 방점을 뒀다. (Javier Andreu-Perez, 2015)

* 1. 관련연구를 통한 고찰

상술한 관련연구를 참고하면서 본 논문은 본 연구의 방향, 전파 방식 등 다양한 방법을 알 수 있었다. 하지만 조류 독감과 빅데이터에 대한 분석 기법, 분석 사례 등을 연구한 논문은 찾을 수 없었고 그나마 빅데이터를 이용해 조류 독감 등 전염병을 예방하는 방법으로는 방역 작업의 세분화, 감염자의 이동 경로 추적 및 분석 등의 이동과 관련된 연구 밖에 없었다.

본 논문은 조류 독감의 전파에 영향을 미치는 요소가 방역 작업과 차량 이동 외에도 환경적 요인이 있을 것이라 확신하고 연구에 임했다. 그러므로 향후 데이터 분석 등을 통해 관련 연구가 미비했던 본 주제에 대한 결론을 도출하여 논문을 작성하고자 한다.

* 제안 작품 소개 (7페이지 이내)

본 논문은 최근 다양한 기법이 발표된 빅데이터 분야를 활용하여 조류 인플루엔자의 발생을 예측하는 것을 주제로 삼고 있다. 제안 배경에서 서술하였듯이 인류의 전염병같은 경우 이미 유엔 산하의 리드컴퍼니에서 빅데이터를 활용해 전염병 발생 예측을 실시하고 있다. 하지만 조류 인플루엔자의 경우 빅데이터를 활용한 전염병 예방 연구 사례가 많지 않아 2016년 11월 발생한 조류 인플루엔자와 같은 엄청난 사회적 손실이 계속해서 발생하고 있다. 이에 본 논문은 조류 인플루엔자로 인한 사회적 손실을 막기 위하여 한 가지 예측 기법을 제시하고자 작성되었다.

본 논문에서 사용할 데이터로는 조류 인플루엔자 발생 지역 및 날짜 데이터와 기온 데이터가 있다. 이들 데이터를 수집, 생성 및 정리한 과정을 서술하고자 한다.

* 1. 데이터 수집, 생성 및 정리

본 논문을 작성하기 위해 먼저 농림축산검역본부 사이트와 기상자료개방포털 사이트로부터 조류 인플루엔자 발생 신고 데이터와 기온 데이터를 얻었다.

조류 인플루엔자 데이터의 경우, 2013년부터 2017년 6월 23일까지 신고 일 단위로 총 4년 간의 데이터가 존재하였다. 하지만 일부 데이터가 너무 파편적이고 결론을 도출하기에는 데이터가 적어서 이를 생략한 2014년 4월부터 2017년 6월 23일까지의 신고 일 단위 데이터를 사용하기로 하였다. 그리하여 농림축산검역본부로부터 조류 인플루엔자 발생 지역, 발생 날짜를 수집하여 데이터 파일에 정리하였다.

기온 데이터의 경우, 기상자료개방포털로부터 조류 인플루엔자 발생 신고 데이터와 상응하는 기온 데이터를 모두 얻었다. 먼저, 1. 기상자료개방포털의 실황자료 페이지에 접속한 이후 조류 인플루엔자 발생 신고 지역을 검색하였다. 그런 다음 2. 발생 신고 날짜를 입력해 본 논문에서 필요로 하는 신고 지역의 당시 기온 데이터를 얻었다. 이 기온 데이터를 기존 작성된 조류 인플루엔자 발생 신고 데이터와 병합, 통합 데이터 파일 초본을 완성하였다.

표 통합된 데이터 파일 예시



데이터 파일 초본을 작성하였으나 본 주제를 진행하는 데에 적합하지 않은 것으로 나타났다. 먼저 본 논문에서는 시각적 그래프와 통계학적 수치 두 가지로 조류 인플루엔자 발생과 기온 간의 상관관계를 얻을 예정이었다. 하지만 논문을 작성하면서 국내 지도 데이터를 pandas 라이브러리에서 구할 수 없었으며, 오로지 미국 지도만이 pandas 라이브러리에 존재하였다. 이럴 경우, 주소 데이터를 지도 상에서가 아닌 문자열만으로 분석해야 할 필요가 있었는데 그래프의 범례를 정하는 것과 그래프 추세선을 그리는 과정에서 문제가 발생하였다. 이에 본 논문에서는 상세한 조류 인플루엔자 발생 지역 주소 정보와 날짜 데이터를 삭제하기로 하였다.

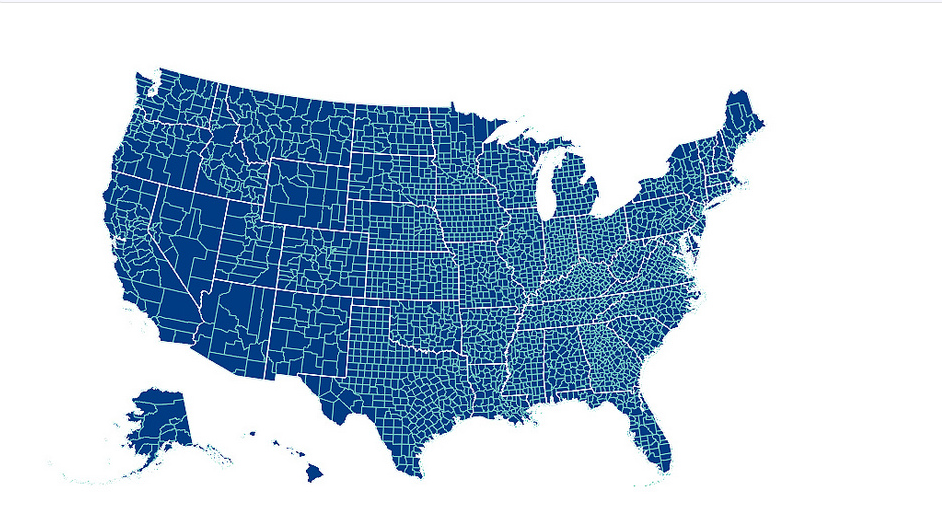


그림 Pandas와 Vincent 라이브러리로 이용할 수 있는 미국 지도 (RobStory, 2013)

또한, 조류 인플루엔자 발생 지역의 파편화가 심해 인접한 지역을 한 구역으로 묶은 뒤, 해당 구역을 1~18, 그리고 20의 숫자로 표현하고 주소 데이터를 대체하기로 하였다. 구역 설정 기준은 아래 표 3과 같다.

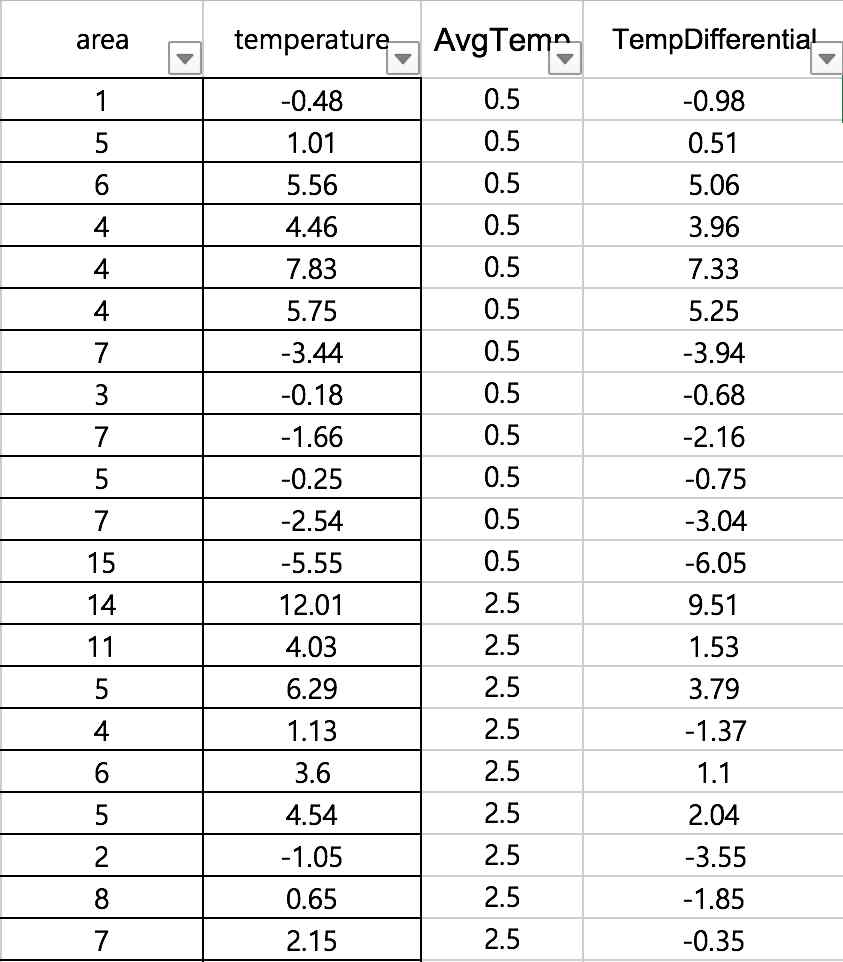
표 조류 인플루엔자 발생 지역을 1~20의 구역으로 나누었다

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| 전남 북부 | 충청 중부 | 경기 남동부 | 전남 남부 | 전북 남부 |
| 6 | 7 | 8 | 9 | 10 |
| 충남 서부 | 경기 서부 | 경기 북동부 | 경기 북서부 | 강원 북서부 |
| 11 | 12 | 13 | 14 | 15 |
| 충북 제천 | 경북 문경 | 경남 남서부 | 부산광역시 | 대구광역시 |
| 16 | 17 | 18 | 19 | 20 |
| 울산광역시 | 전남 보성군 | 전남 동부 | - | 제주도 지역 |

구역 설정 이후 기존 데이터 파일에서 주소 문자열 데이터를 구역 숫자 데이터로 대체하였다. 또한, 해당 지역의 기온 데이터, 즉 절대적인 수치보다는 해당 지역과 전국 평균 기온과의 차라는 상대적인 수치를 이용하는 것이 상관관계를 분석하는 데에 더 효과적이라고 판단하였다. 이에 조류 인플루엔자 발생 신고 당시의 전국 평균 기온 데이터도 기상자료개방포털로부터 얻은 후 데이터 파일에 삽입, 조류 인플루엔자 발생 당시 기온과 전국 평균 기온 간의 차도 데이터 파일에 삽입하였다. 이렇게 수정된 데이터 파일이 완성되었다. 데이터 파일의 속성값은 구역(*area*), 해당 지역의 기온(*temperature*), 당시 전국 평균 기온(*AvgTemp*) 그리고 발생 지역 기온과 전국 평균 기온의 차(*TempDifferential*)가 있다.

표 4는 수정된 데이터 파일의 예시이다.

표 수정된 통합 데이터 파일 예시



논문에 사용할 통합 데이터 파일이 작성된 후, 이를 이용해 선 그래프와 회귀분석 기법을 이용하여 연관성을 얻는다. 이는 다음 장인 구현 및 결과분석에서 다루고자 한다.

* 1. 데이터 분석 툴 및 기법 선택 과정

데이터 수집이 완료된 후, Python 언어를 통해 데이터를 분석하고자 하였다. 처음 분석 라이브러리로 Pyhthon의 altair 라이브러리를 선택하였다. 인터넷 상의 altair 라이브러리와 관련한 예시가 많아 그림 3과 같은 그래프를 작성하는 데에는 어려움이 없었다.

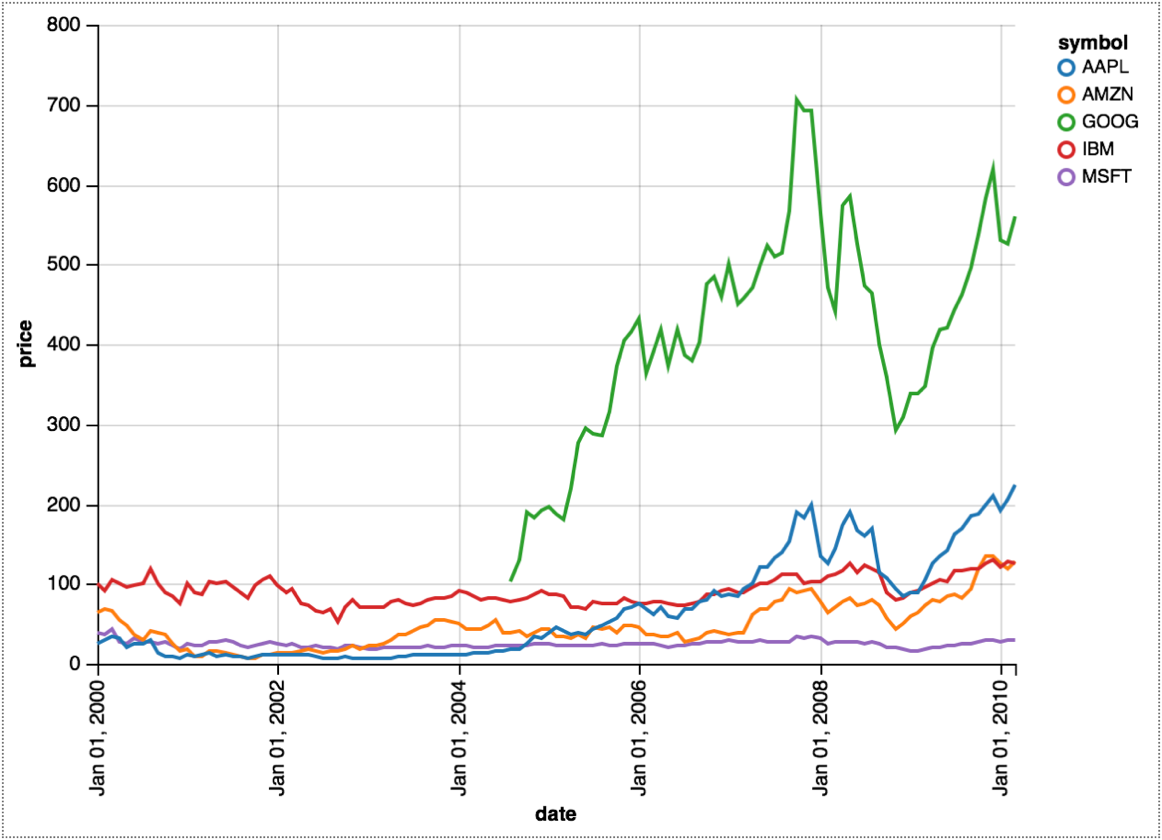


그림 Python 언어의 altair 라이브러리를 이용해 작성한 예시 그래프(코드 부록 참조)

하지만 정작 사전에 작성된 통합 데이터 파일을 불러오는 데 문제가 발생하였다. Csv로 작성된 데이터 파일을 altair 라이브러리를 통해 불러오는 데 계속해서 실행이 되지 않은 것이다. 해결책을 찾기 위해 altair 라이브러리 홈페이지에서 수많은 자료를 검색하였지만 파일을 불러오는 동작이 끝내 실행되지 않았고, 결국 다른 라이브러리를 사용하기로 결정하였다.

Python 언어 라이브러리 중 pandas를 이용해 그래프를 작성하는 것이 편리하다고 판단하여 pandas 라이브러리를 이용하기로 결정하였다. 통합 데이터 파일인 csv 파일 불러오기까지 성공하여 그래프를 작성할 수 있었다.

데이터 분석 기법에서는 인터넷을 통해 데이터 분석 기법 종류를 검색하였다. 이 중 본 논문 주제에 적합한 분석 기법 몇 가지를 정리하였다. 아래와 같다.

1. 결정트리
2. 회귀분석
3. 연관규칙

이렇게 세 가지 기법으로 정리하였다. 이 중 결정트리의 경우, 본 논문 주제의 데이터 중 기온 데이터를 특정 방향으로 선택할 수 있는 기존이 존재하지 않아 선택할 수 없었다. 회귀분석의 경우, 조류 인플루엔자 발생 신고 데이터와 기온 데이터 간의 상관관계를 수치적으로 표현할 수 있으므로 분석 기법으로 채택하였다. 연관규칙의 경우, 여러 item set을 통해 다른 종류의 set을 도출할 수 있다는 것을 설명하는 규칙이지만 본 논문에서는 조류 인플루엔자 발생 신고 데이터와 기온 데이터 단 두 종류의 데이터 set만이 존재한다. 그러므로 연관규칙 기법 또한 본 논문 주제와는 맞지 않아 최종적으로 분석 기법으로는 회귀분석만을 선택하였다.

■ 구현 및 결과분석 (7페이지 이내)

본 논문에서 구현하여야 할 요소로는 시각적 분석을 위한 선 그래프와 통계학적 분석을 위한 회귀분석 기법이 있다.

그래프 구현의 경우, altair 라이브러리를 이용했으나 파일 불러오기 문제로 진척이 없어 결국 pandas 라이브러리를 이용하기로 결정하였다. Python 개발 환경으로는, 컴퓨팅 운영체제 Mac OS X, Safari 브라우저를 이용해 Anaconda를 구동, Jupyter Notebook을 이용하여 python 버전 2.7.14를 이용하였다. 초기에는 python 3.6.3을 이용하였으나 라이브러리와 csv 파일 불러오는 법 등이 인터넷에 자세히 나와있지 않아, 가장 많이 사용되고 있는 2.7.14 버전으로 변경하였다. 이후, pandas 라이브러리를 익혀가면서 몇 가지 예시 그래프를 작성하기 시작했고, 본 논문이 얻고자 하는 결론을 가장 잘 나타내는 그래프 작성 코드를 선택하여 그래프를 작성하였다.

다음은 그래프를 그리기 위해 이용한 csv 파일 캡처본이다.

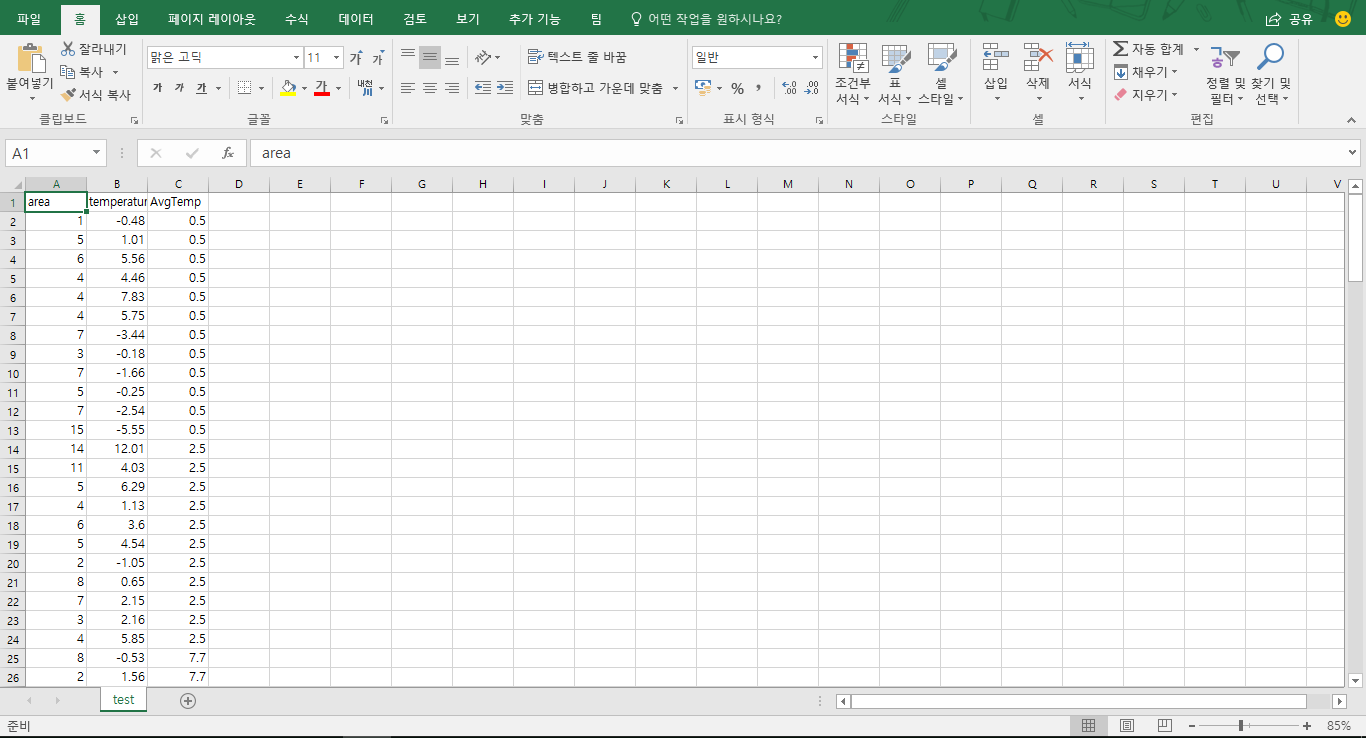


그림 선 그래프를 작성하기 위해 사용한 csv 파일

그래프를 그리기 위한 코드는 부록에 작성하였다. (Sanchez, 2016)

기존 통합 데이터 파일에는 전국 평균 기온과 조류 인플루엔자 발생 지역 기온의 차인 기온 차(*TempDifferential*) 속성으로 존재하였다. 본 그래프에서도 해당 데이터를 사용할지 고민하였으나, 조류 인플루엔자 발생 지역 기온과 전국 평균 기온의 추세선을 분석하며 더욱 자세란 분석이 자세할 것이라 판단하여, 해당 csv 파일에는 기온 차 속성이 빠져있다.

그림 5에서 파란색 선은 1부터 20까지 해당되는 구역을 의미하는 추세선을 의미한다. 연두색 선은 조류 인플루엔자 발생 지역의 기온을 나타내는 추세선이며, 빨간색 선은 조류 인플루엔자 발생 당시 전국 평균 기온을 나타내는 추세선이다.

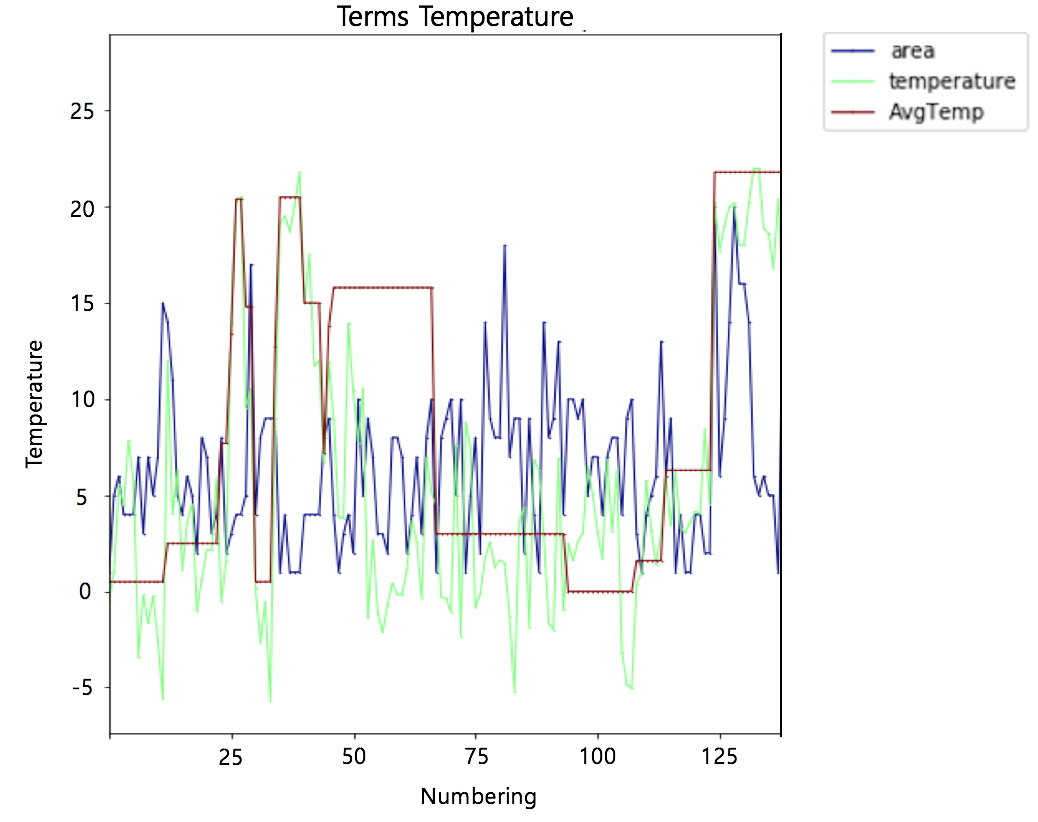


그림 구역, 기온 그리고 평균 기온 데이터를 기반으로 작성한 line graph(SanchezJaime, 2016)

그림 5의 선 그래프에서는 조류 인플루엔자 발생 지역과 기온 및 전국 평균 기온 간의 명확한 연관성은 얻을 수 없었다. 그리하여 결국 조류 인플루엔자 발생 지역과 기온 차 데이터만으로 작성된 그래프를 다시 완성하였다.

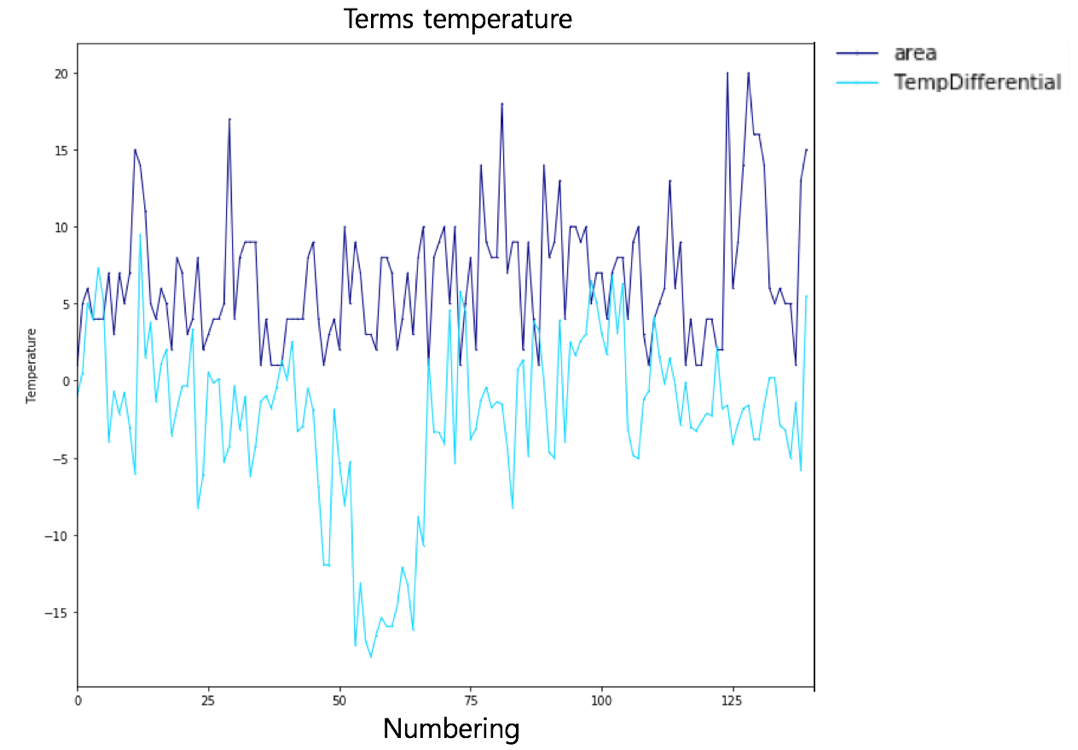


그림 구역, 기온 차만을 이용해 작성한 line graph

조류 인플루엔자 발생 신고 지역과 기온 차만을 이용해 선 그래프를 작성하였지만 이 역시 명확한 결론을 도출할 수는 없었다. 그리하여 그래프를 이용한 결론 도출 보다는 회귀분석 기법을 이용해 객관적인 수치로 결론을 증명하고자 하였다.

회귀분석 기법을 이용하기 위해서는 독립변수와 종속변수를 설정할 필요가 있다. 이에 독립변수는 조류 인플루엔자 발생 신고 구역 번호, 종속변수는 신고 지역 기온와 전국 평균 기온 간의 차로 설정하였다.

독립변수 하나와 종속변수 하나, 총 두 개의 요소가 존재하는 회귀분석을 단순선형 회귀분석이라 한다. 이를 계산하는 식은 아래와 같다.

여기서 y는 독립변수, 즉 본 논문에서는 조류 인플루엔자 발생 신고 구역 번호이다. X는 종속변수인 기온 차를 나타내며 과 는 모두 회귀계수이다. 이 두 계수는 추정을 통해서만 얻을 수 있다. 는 임의의 수이다.

수학 공식을 통해 결론을 도출하고자 하였으나 방대한 데이터와 수학적인 어려움으로 본 논문에서는 엑셀의 회귀분석 수식을 이용하고자 하였다. 표 5는 엑셀 파일을 통해 얻은 조류 인플루엔자 발생 구역 번호와 기온 차의 회귀분석 결과표이다.

표 엑셀 파일을 통해 얻은 결과표

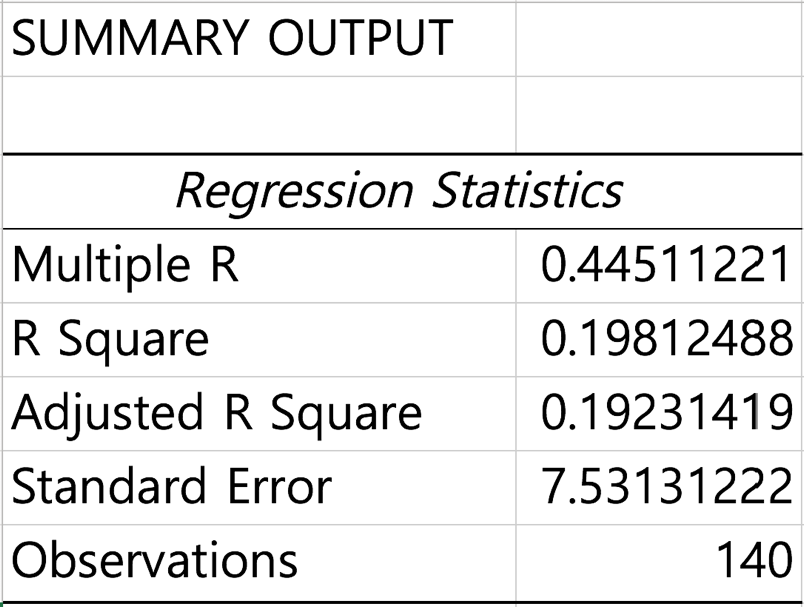


표 5에서 Multiple R은 독립변수와 종속변수 간의 연관성을 나타낸다. 즉, 조류 인플루엔자 발생과 기온 차 간에는 44.5%의 연관이 있다. 한 요소가 다른 요소에 의해 절반에 가까운 영향을 받는 다면 이는 적지 않은 연관성이 있다는 것을 의미한다. 예를 들어, 대부분 사람이 암에 걸리면 사망할 가능성 이 매우 높다고 생각하지만, 실제 암에 의한 사망률은 27.8% 에 불과하다 (국가암정보센터, 2017). 즉, 암과 사망률 간의 연관성은 27.8에 불과 한 것이다. 이렇듯 회귀분석을 이용해서 얻은 객관적 수치를 통해 조류 인플루엔자 바이러스와 기온 간의 상관관계가 높음을 알 수 있다.

기타 분석 기법을 이용해 또다른 객관적인 수치를 얻으려 하였으나 본 논문 주제와 맞는 분석 기법을 찾지 못하였다. 이에 선 그래프를 통한 시각적 자료와 회귀분석 기법을 통한 통계학적 분석으로 결론을 맺고자 한다.

■ 결론 및 소감 (2페이지 이내)

Python의 pandas 라이브러리를 이용해 작성한 선 그래프, 즉 시각적 데이터로는 본 논문의 결론이 명확한지는 알 수 없었다. 하지만 추세선의 움직임을 통해 어느 정도의 연관성이 있음을 알 수 있었다. 객관적인 수치를 얻기 위해 회귀분석 기법을 이용하였고 이를 통해 조류 인플루엔자 발생과 기온 간 44.5%의 연관성을 갖는 다는 결론을 얻어 그 연관성이 크다는 것을 결론으로 도출할 수 있었다.

본 논문 주제와 관련된 다른 논문 수가 적다는 사실은 이 주제 자체가 신선하고 아직 많은 연구가 진행되지 않았으므로 새로운 시도를 하고 있다는 것이라 생각한다. 인간의 전염병을 예방, 예측하기 위해 빅데이터를 이용한 사례는 이미 많이 존재한다. 통신사와 세계 여러 기관이 빅데이터를 제공하면서 인류 전염병 예방, 예측도 4차 산업혁명에 따라 그 방식이 진화하고 있는 것이다. 하지만 조류 인플루엔자의 경우 아직도 구시대적인 방식, 즉 방역 조치 등의 옛 방식에 의존하고 있는 실정이다. 관련 연구도 많지 않아 빅데이터를 활용한 조류 인플루엔자 예방, 예측 방식이 활용되기까지는 아직도 많은 시간이 남아 있다고 생각한다. NDSL이나 RISS에서 조류 인플루엔자와 빅데이터와 관련된 논문을 검색했으나 단 몇 가지의 논문만이 검색되었고 이조차도 예방을 위한 목적이 아닌 “조류 인플루엔자에 대한 인식과 SNS 간의 관계”와 같은 주제만 있었다. 그러므로 본 논문의 주제를 선정한 것에 큰 보람을 느꼈다.

아직은 시각적 데이터가 부족하고 통계학적 자료도 회귀분석 기법 단 하나만 이용하여 향후 결론을 보완할 여지는 많다고 생각한다. 또한, 논문 작성 초반에는 기온 뿐만 아니라 주변 환경 요소, 예를 들어 습도, 날씨 등 다양한 데이터를 기반으로 조류 인플루엔자와의 연관성을 연구하려 했었다. 하지만 데이터를 수집하는 데에 한계가 있고 양 또한 너무 방대해져 분석하는 것에 어려움이 있을 것이라 판단하였다. 그리하여 주변 환경 요소 중 기온 데이터 하나만을 이용해 종속변수로써 사용하였다. 이후 다양한 데이터와 분석 기법을 이용해 본 논문의 결론을 보완할 것이며 빅데이터를 이용하여 조류 인플루엔자를 막는 데에 보탬이 되고자 한다.

본 논문 작성을 진행하면서 처음으로 Python 언어를 다뤘다. 이를 위해 여름 방학 기간 동안 Python 언어를 Codecademy 사이트[[2]](#footnote-2)를 통해 독학했으며 이는 Github에서 확인할 수 있다. 그렇게 배운 Python 언어로 그래프를 그리기 위해 수많은 웹사이트를 참고하였다. 처음에는 Pandas 라이브러리가 아닌 Vincent와 Vega를 사용하여 그래프를 그리려 했으나 파일을 불러오는 것에 실패하였다. Vincent와 Vega에 더욱 많은 그래프 예시가 있었으나 파일 불러오기 및 그래프 내 범례 수정하기 등 여러 어려움이 있어 다른 라이브러리인 pandas를 이용해 선 그래프를 작성하기로 최종 결정하였다. 그렇게 선 그래프를 완성할 수 있었다.

그런 다음 회귀분석 기법도 사용하여 객관적인 수치를 제시하여 논문에서 원하던 결론을 얻을 수 있었다.

# 참고문헌

[1] AugspurgerTom. (2016년 4월 28일). “Modern Pandas (Part 6): Visualization”. datas-frame:

[2] CompactNations GlobalUnited. (2017). “2017 United Nations Global Compact Progress Report: Business Solutions to Sustainable Development”. United Nations Global Compact:

[3] FriendR. “[Python pandas] text, csv 파일 불러오기 : pd.read\_csv() ”. R, Python 분석과 프로그래밍: http://rfriend.tistory.com/250

[4] Jnana. (2017년 1월 11일). “2016 - 2017년 겨울 AI(조류 독감) 발생지역 (1/11 갱신).” 티스토리: http://jnana.tistory.com/818

[5]K.SeppanenJouni. (2009년 5월 12일). “How do you change the size of figures drawn with matplotlib? ”. Stackoverflow: https://stackoverflow.com/questions/332289/how-do-you-change-the-size-of-figures-drawn-with-matplotlib

[6] PandasOrganization. “Visualization”. Pandas documentation: https://pandas.pydata.org/pandas-docs/version/0.18.1/visualization.html

[7] PinkWink. (2015년 8월 6일). “Python Pandas를 이용해 분석한 데이터를 그래프로 표현하기... ”. PinkWink: http://pinkwink.kr/773 [PinkWink]

[8] rubennj. (2015년 3월 1일). “.plot() method slow compared to Matplotlib”. Github: pandas-dev page: https://github.com/pandas-dev/pandas/issues/9567

[9] SanchezJaime. (2016년 2월 2일). “Data Visualization with Plotly and Pandas”. SODA Developers: https://dev.socrata.com/blog/2016/02/02/plotly-pandas.html

[10] SomaJonathan. “Labeling your axes in pandas and matplotlib”. Lede: http://jonathansoma.com/lede/data-studio/matplotlib/labeling-your-axes-in-pandas-and-matplotlib/

[11] TeamCommunicationPlot.ly. “ Cufflinks in Python”. Plot.ly : https://plot.ly/ipython-notebooks/cufflinks/

[12] TeamCommunicationPlotly. “Lines on Maps in Pandas”. Plotly: https://plot.ly/pandas/lines-on-maps/

[13] TeamCommunicationPlotly. “Map Subplots and Small Multiples in Pandas”. Plotly: https://plot.ly/pandas/map-subplots-and-small-multiples/

[14] wrobstory. “Mapping Data in Python with Pandas and Vincent”. Pelican blog: http://wrobstory.github.io/2013/10/mapping-data-python.html에서 검색됨

[15]기상청.“실황분석자료”.기상자료개방포https://data.kma.go.kr/data/rmt/rmtList.do?code=400&pgmNo=570

[16] 김재만. (2015년 11월). 조류인플루엔자 발병지역을 예상하는 ‘빅데이터’. “MSIT Webzine”.

[17] 김종현(KimHyun) ,송태준(Song,Tae Joon) ,김상수(Kim,Sang Su) ,안동혁(Ahn,Dong Hyuk)Jong. (2014). 빅데이터 기반 실시간 농축산 검역을 위한 확산경로 탐색과 역학조사 모델링 연구. “한국IT서비스학회 학술대회 논문집” (페이지: 5-6). 서울: 한국IT서비스학회.

[18] 이주희. (2017). “가축방역 자료실”. 농림축산검역본부:

[19] 정빛나. (2017년 1월 5일). “AI 재앙 언제 끝나나…"피해규모 1조원"”. 연합뉴스: http://www.yonhapnews.co.kr/bulletin/2017/01/04/0200000000AKR20170104182700030.HTML

[20] 정영철배용근최익수,. (2016). 사회안전을 위한 빅데이터 활용의 재난대응 정책. “한국정보통신학회논문지” (페이지: 684-685). 서울: 한국정보통신학회.

[21] 허지윤. (2017). 페이스북 이용을 통한 정보획득과 사회자본이 질병 예방행동의도에 미치는 영향. “연세대학교 석사학위논문” (페이지: 4). 서울시: 연세대학교

■ 부록

1. Altair를 이용해 작성한 예시 그래프

from altair import \*

Chart(Data(

format=DataFormat(

type='csv',

),

url='https://vega.github.io/vega-lite/data/stocks.csv',

),

description='Stock prices of 5 Tech Companies Over Time.',

).mark\_line().encode(

color='symbol:N',

x='date:T',

y='price:Q',

)

2. Pandas 라이브러리를 이용해 그린 선 그래프 코드

그래프 생성에 작정된 코드는 다음과 같다.

import pandas as pd

import numpy as np

import matplotlib

import cufflinks as cf

import matplotlib.pyplot as plt

import plotly

import plotly.offline as py

import plotly.graph\_objs as go

from pylab import rcParams

cf.go\_offline()

py.init\_notebook\_mode()

data = pd.read\_csv('/Users/gundu/Desktop/test.csv')

plot = data.plot(title='Temrs temperature', lw=1, colormap='jet', marker='.', markersize=1)

plot.set\_xlabel("numbering")

plot.set\_ylabel("Temperature")

rcParams['figure.figsize'] = 16, 10

plt.show()

1. 데이터 출처: http://www.bloter.net/archives/249712, https://meraki.cisco.com/blog/2013/02/cisco-forecasts-global-mobile-traffic/ [↑](#footnote-ref-1)
2. https://www.codecademy.com [↑](#footnote-ref-2)