

V

https://github.com/dDragun/FinalProject

주변 환경 분석을 통한 조류 인플루엔자 바이러스 감염 경로 예측 모델

2

7xq

1. 요약

작년 11월부터 시작된 조류 독감으로 인해 대한민국은 산란계의 떼죽음으로 전례 없는 계란 값 폭등을 겪었다. 한국 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. 사실 기온 외에도 습도, 강수량 등 외부 환경 요인을 늘릴 수 있으나 본 논문에서는 기온과의 관계만 파해지고자 한다. Regression Analysis 기법, Association Rule Learning 기법 등을 이용해 조류 독감과 외부 기온의 상관관계를 파악한 후 조류 인플루엔자 바이러스의 경로를 예측하고자 한다. 조류 독감 발생 데이터는 농림축산검역본부 사이트에서 공개 데이터로 얻을 수 있고 기온 등은 기상청 데이터를 통해 얻을 수 있다. 본 논문을 통해 조류 독감의 예상 전파 지역을 미리 파악한 후 대처하여 조류 독감 발생 빈도 감소를 기대해볼 수 있다.

1. 서론

작년 11월부터 발생하기 시작한 조류 독감은 대한민국 사회에 지대한 영향을 끼쳤다. 이전부터 조류 독감으로 인한 피해는 있었으나 그 여파가 일반 소비자 식탁에까지 미친 건 이번이 처음이다. 계란 값은 며칠만에 두 배 이상으로 치솟았으며 계란을 이용한 음식이 많은 한국인의 밥상 특성상 그 영향은 이루 말할 수 없을 정도로 지대했다. 금년 3월 기준 살처분된 가금류의 수는 3000만 마리 이상으로 기존 사육되는 가금류 수의 20%를 차지한다. (정빛나, 2017) 이렇듯 조류 독감으로 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 독감으로 인해 대량으로 살처분 되었고, 계란은 팔리지 않으니 양계업주들은 그야말로 사면초가에 빠진 것이다.

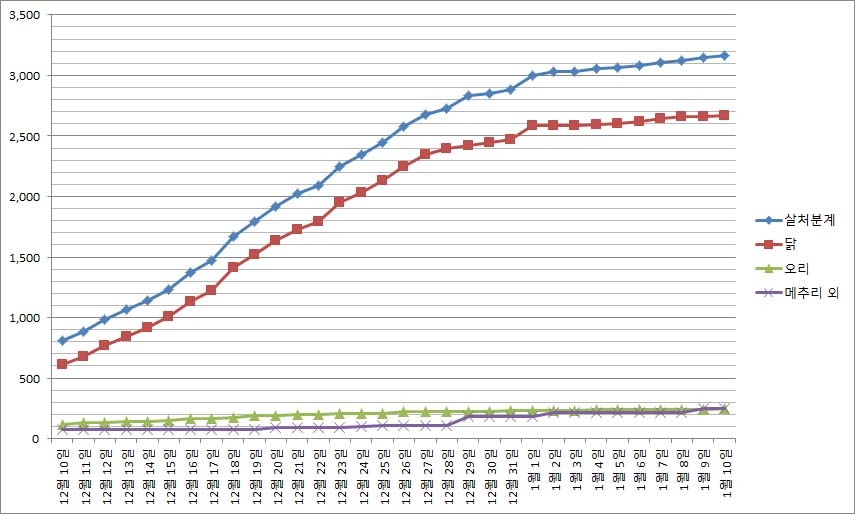


Figure 1금년 1월10일까지의 살처분계 상황 그래프

이렇듯 조류 독감의 영향은 10개월이 지난 지금까지도 소비자와 양계업주 모두에게 영향을 끼치고 있다. 살충제 파동으로 계란의 소비가 급락하자 계란 가격은 다시 이전 또는 그보다 더 낮아졌다. 하지만 건강에 대한 인식이 매우 커진 현 상황에서 살충제 사용으로 계란에 대한 이미지가 나빠진 현 상황에서 계란 소비의 증대 기대를 하기는 어려울 것으로 보인다.

사회적으로나 경제적으로나 한국 사회에 심각한 피해를 준 조류 독감을 예방하는 방법으로는 방역 작업이 전부일 정도로 기술적 지원이 안 되고 있는 상황이다. 하지만 UN(이하 유엔) 산하의 LEAD Company(이하 리드 컴퍼니) 사례를 통해 조류 독감을 빅데이터 분석을 통해 예방할 수 있을 것이라 생각한다. 리드 컴퍼니의 주 목적은 빅데이터로 감염병의 전파 경로를 예측해 이를 사전에 예방하는 것이다. 이 방법을 조류 독감에 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 독감의 예상 전파 경로를 예측해 감염을 방지하고, 사회적 혼란을 막자는 데에 있다.

본 논문의 최종 목표는 조류 독감을 1차 발생에 그치게 하는 것이다. 사실 조류 독감은 철새에 의해 유입되므로 완전 예방은 불가능하다. 하지만 바이러스의 전파 경로를 예측해 막는다면 조류 독감으로 인한 피해를 대폭 줄일 수 있을 것이다. (김재만, 2015)

본 논문은 크게

1. 데이터 추출 및 정리 상황 요약
2. 데이터를 이용한 그래프 추출
3. 그래프를 기반으로 해당 년, 월의 평균 기온과 조류 독감 발생지역의 기온을 분석하여 그 연관성을 추출
4. 결론

이 4가지 부분으로 진행될 예정이다.

데이터 추출 및 정리 상황 요약에서는, 조류 독감 신고 상황을 농림축산검역본부 사이트에서, 조류 독감 발생 지역의 기온 데이터 및 당시 한국의 평균 기온 데이터는 기상청 사이트에서 데이터를 얻을 수 있었다. 그런 후 조류 독감 발생 지역과 날짜, 그리고 당시 해당 지역 기온 및 당시의 한국 평균 기온 데이터를 병합해 정리하였다.

데이터를 이용한 그래프 추출에서는, Python의 Pandas 등의 라이브러리를 이용해 그래프를 얻을 수 있었다. 그래프 추출 시작 단계에서 Python의 Altair 라이브러리를 이용해 그래프를 작성할 계획이었으나 Pandas의 사용법 및 사례가 더 많다는 것을 발견하여 Pandas를 이용하기로 최종 결정하였다. (PinkWink, 2015)

그래프를 기반으로 해당 년, 월의 평균 기온과 조류 독감 발생지역의 기온을 분석하여 그 연관성을 추출하는 단계는 아직 진행 중이지만 분석 작업을 단시간 내에 끝마칠 수 있을 것이다.

분석이 끝난 후 결론으로 본 논문을 끝마칠 계획이다.

분석 단계를 끝마친 후 최종보고서를 작성하여 제출할 계획이다. 중간 보고서를 작성하고 있는 현재까지도 그래프 분석은 진행되고 있으며 진전이 있을 때마다Github에 업로드 할 계획이다.

1. 관련연구
   1. 빅데이터와 전염병의 연관성

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 되었지만 구체적으로 조류 독감의 확산과 빅데이터와 관련된 연구가 진행된 경우가 매우 드물다. 실제로 RISS와 NDSL에서 “조류 독감 빅데이터” 키워드로 검색할 시 RISS에서는 관련 논문 하나, NDSL에서는 아예 관련 논문이 검색되지 않는다. RISS에서 검색된 논문은 「빅데이터 기반 실시간 농축산 검역을 위한 확산경로 탐색과 역학조사 모델링 연구」[1]으로 2010년 발생한 구제역과 농장간 차량 이동을 연관 지어 특정 차량 방문 시 바이러스가 전염된다는 주제로 작성되어 있다. 이 논문에 의하면, 기존 역학 조사는 특정 농장 방문 차량 리스트를 추출하거나 특정 차량의 방문 농장 리스트를 추출하는 기법을 사용하고 있으나 이를 특정 농장 방문 차량을 유형별, 목적별, 기간별 등으로 분석하고 특정 차량의 방문 농장을 유형별, 가축종류별, 기간별, 발병이력여부별 등으로 분석하는, 보다 자세하고 빅데이터를 이용해 방대한 데이터를 통해 유의미한 데이터를 추출하는 것에 방점을 두고 기존 시스템을 개선하고자 하는 이론을 제시하였다. (김종현(KimJong, 2014)

* 1. 효과적인 논문 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. 논문[2]를 통해 본 학생은 본 논문의 주장을 매스 미디어 등의 권위 있는 언론매체가 보도해야만 실질적인 효과가 있으리라 생각한다.

* 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) 이를 통해 작년 11월부터 발생한 조류 독감 피해는 정부의 미흡한 대응 방식도 한 몫한다는 것을 얘기하고 있다. 만약 정부가 조류 독감을 빅데이터를 이용해 적극적으로 예방한다면 서론에서 서술한 것과 같은 재앙이 발생하지 않았을 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 사항이라하고 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 학생은 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 학생은 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.

1. 제안 작품 소개 (4페이지 이내)

본 논문은 데이터를 직접 추출, 정리하여 하나의 엑셀 파일로 데이터를 병합하였다. 조류 독감 신고 데이터는 2014년부터 2017년까지의 데이터를 이용하였다.

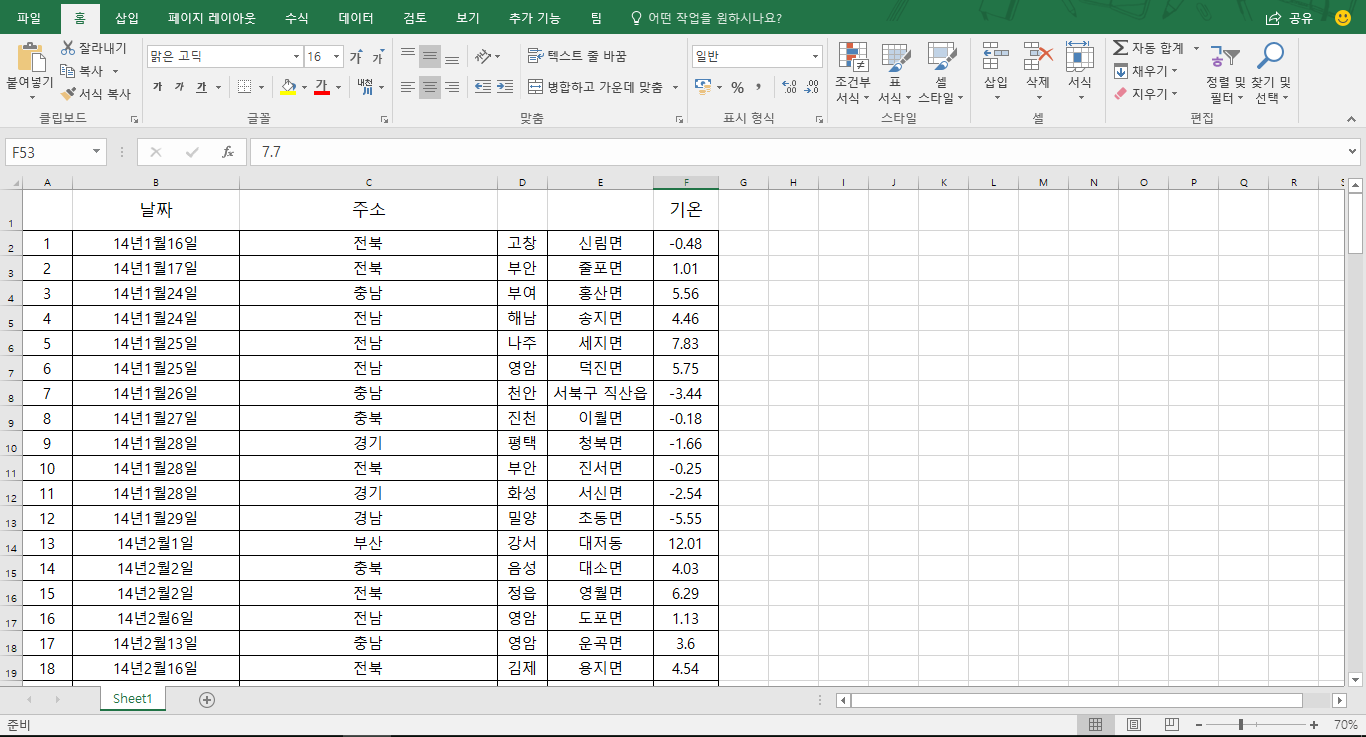


Figure 2 통합된 날짜-주소-기온 데이터 엑셀 파일 (기상청, 날짜 정보 없음)

위 이미지에서 날짜는 조류 독감 발생 신고가 들어온 날짜이고 주소 (이주희, 2017)는 발생 지역 주소를 의미한다. 기온은 해당 지역의 당시 기온을 나타내고 있다. 여기서 본 학생은 국내 지도 데이터를 이용하여 위 주소를 지도 상에서 자동으로 나타내고, 기온 또한 표기하여 감염 경로를 알아내고자 했다. 하지만 Python 관련 인터넷 페이지에서는 미국 지도 데이터만 있을 뿐 국내 지도 데이터는 찾을 수 없었다. (Team P. C., Lines on Maps in Pandas, n.d.) (PandasOrganization, n.d.) (Team P. C., Map Subplots and Small Multiples in Pandas, n.d.)

그래서 결국 지도 상으로 주소와 기온 데이터를 삽입하는 방법은 포기하였고 다른 방법을 고안하기로 했다. 일단 아래 이미지를 살펴보자.

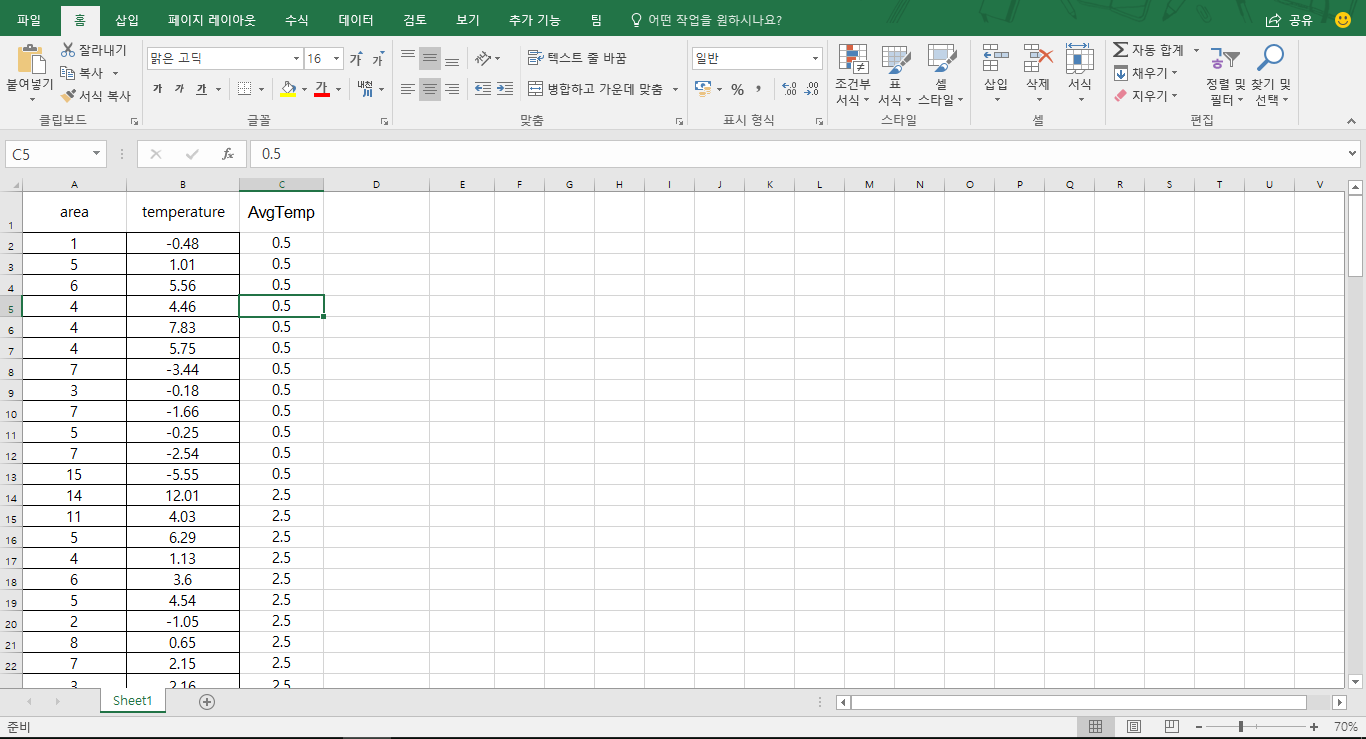


Figure 3 정리된 날짜-지역-기온-한국 평균기온 데이터 엑셀 파일

위 이미지에서 area같은 경우 본 학생이 조류 독감 발생 지역을 몇 개의 구역으로 나눠 분류한 결과물이다. 1번 구역부터 2 (FriendR, 날짜 정보 없음)0번 구역까지 있으며 1번부터 19번까지는 반도 내 구역, 20번은 제주도를 의미한다. temperature는 신고가 들어온 지역의 당시 기온, AvgTemp는 당시 대한민국의 평균 기온을 의미한다. 해당 데이터는 1번을 시작으로 조류 독감 신고가 들어온 순서대로 나열한 데이터이다.

이 데이터를 이용해 어떤 식으로 분석해야 하는지는 아직도 고민 중이다. 우선 현재까지 생각해낸 방법은, area 데이터와 temperature, AvgTemp 데이터를 모두 꺾은 선형 차트로 만들어 평년 기온과 당시 해당 구역 기온을 비교해 기온과 바이러스 전파 경로의 상관 관계를 파악할 예정이다. (AugspurgerTom, 2016)

1. 구현 및 결과분석 (4페이지 이내)

그래프 구현은 Python의 Pandas 라이브러리를 이용한다. Python 개발 환경은 Anaconda를 이용한 Jupyter Notebook을 이용, Python 버전은 Python 3.5.1이다. Jupyter Notebook은 구글 Chorme 브라우저를 이용해 사용했다.

다음은 그래프를 그리기 위해 이용한 csv 파일 캡처본이다.

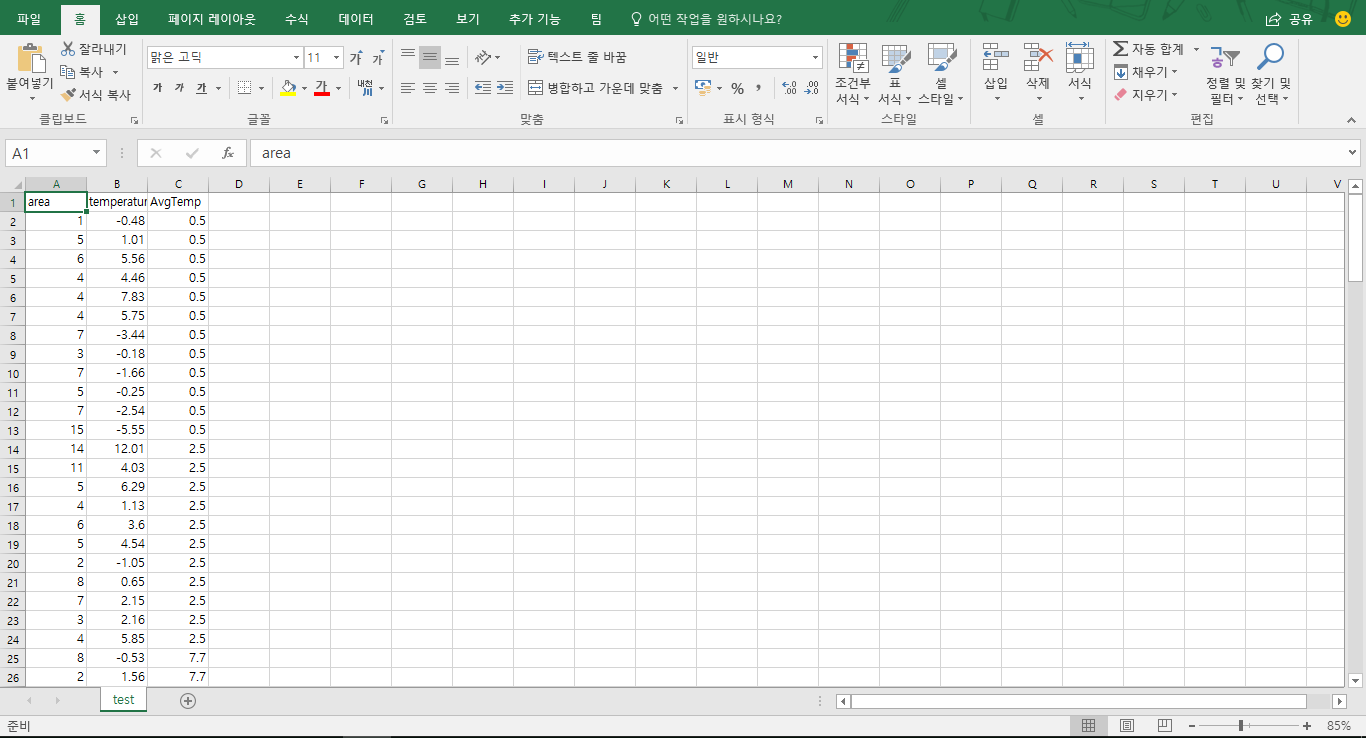


Figure 4 그래프를 그리기 위해 사용한 csv 파일, Figure 5를 csv 파일로 변환만 했다

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

위 코드에서 data는 csv 파일을 불러오고 이를 저장하는 변수이다. 또한 data.plot를 이용해 그래프를 만들었으며 plot.set\_xlable과 plot.set\_ylable을 이용해 그래프를 그렸다. 그래프 사이즈는 width 16 inch, height 10 inch이다. (rubennj, 2015)

이렇게 만든 그래프를 plt.show()로 나타내며 아래 이미지와 같다. (SomaJonathan, 날짜 정보 없음)

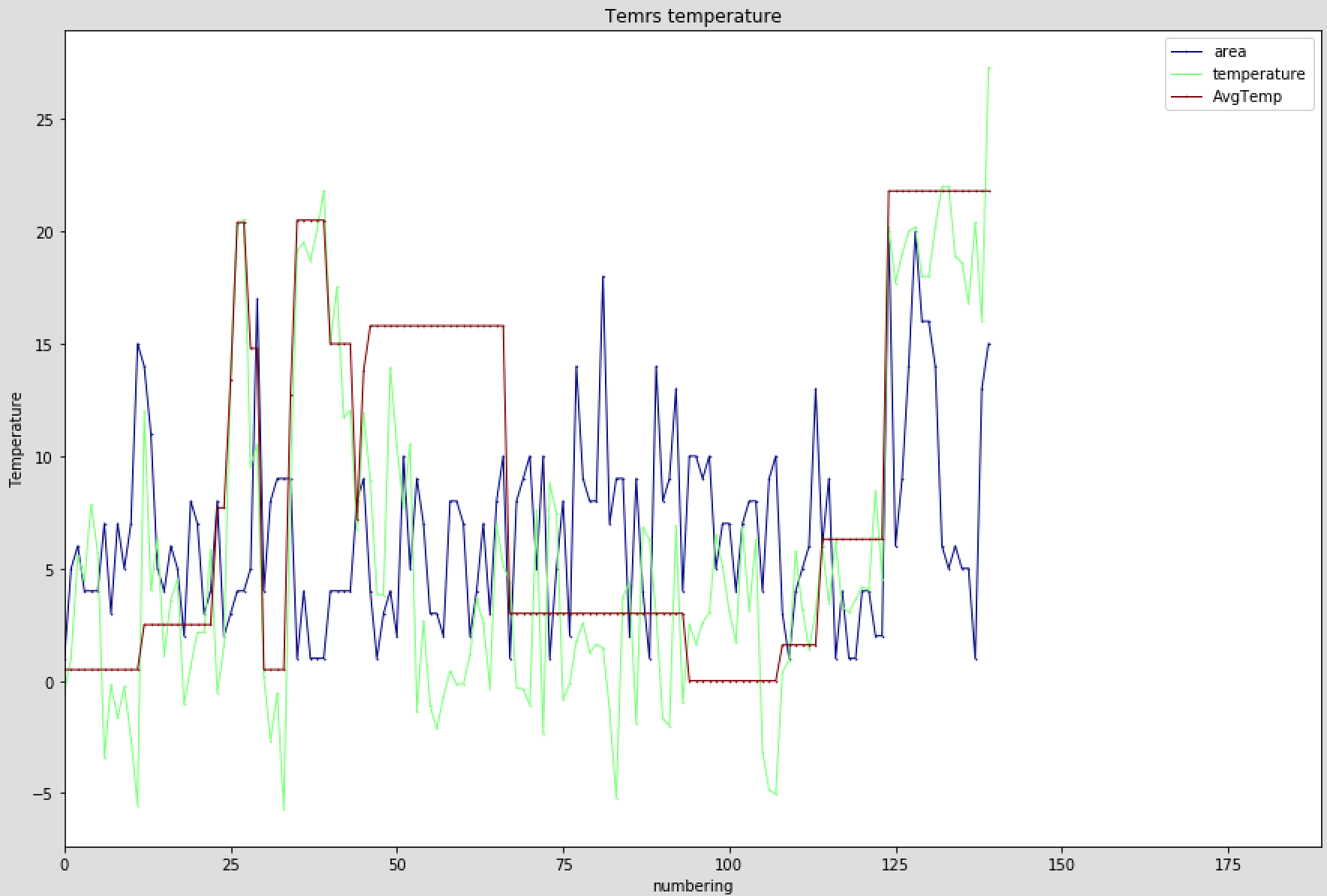


Figure 5 구역, 기온 그리고 평균 기온 데이터를 기반으로 작성한 line graph (K.Seppanen, 2009)

위 이미지에서 area는 높낮이를 보는 것이 아닌 높이가 달라질 경우 해당 지역 기온와 한국 평균 기온의 상관 관계를 보면 된다. 초록색 temperature선과 갈색 AvgTemp선을 비교하여 그 상관 관계를 추후 도출할 예정이다. (Team P. C., n.d.)

1. 결론 및 소감

본 논문 주제와 관련된 다른 논문 수가 적다는 사실은 이 주제 자체가 신선하고 아직 많은 연구가 진행되지 않았으므로 본 학생이 새로운 시도를 하고 있다는 것이라 생각한다. 앞으로 분석을 진행하면서 새로운 현상을 찾을 것이고 이를 최종 발표회에서 발표할 것이다.

본 논문 작성을 진행하면서 처음으로 Python 언어를 다뤘다. 이를 위해 여름 방학 기간 동안 PyThon 언어를 Codecademy 사이트를 통해 독학했으며 이는 본 학생의 Github에서 확인할 수 있다. 그렇게 배운 Python 언어로 그래프를 그리기 위해 정말 많은 웹사이트를 참고하였다. 처음에는 Pandas 라이브러리가 아닌 Vincent와 Vega를 사용하여 그래프를 그리려 했으나 웹사이트에서는 Pandas를 이용한 경우가 더 많아 본 학생도 Pandas 라이브러리를 이용해 그래프를 작성하게 되었다.

아직 기온과 조류 독감 바이러스의 전파와 관련된 구체적인 상관 관계가 나온 것은 아니지만 현재 본 학생이 그래프를 분석한 바로는 일정한 관계가 있을 것으로 예상한다. 구체적인 분석 기법 및 결론은 최종 발표회에서 다룰 예정이다.

만약 기온과 조류 독감 바이러스의 상관 관계가 명확히 밝혀지면, 이 이론을 다른 전염병 바이러스에도 대입해 그 전파 경로를 예측하고 미리 예방할 수 있을 것이다.

1. 부록:

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()

# 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