기온 분석을 통한 조류 인플루엔자 바이러스 감염 발생 예측 모델

두건(*), 오수진(**), 김응모(***)
(*) 성균관대학교 소프트웨어대학, dayikagun34@naver.com
(**) 성균관대학교 정보통신대학, bgbanana4@gmail.com
(***) 성균관대학교 소프트웨어대학, ukim@skku.edu

A model for predicting the avian flu virus infection through temperature analysis

Chien Tu(*), Oh Su-jin(**), Kim Ung-Mo(***)
(*) Sung Kyun Kwan University, College of Software
(**) Sung Kyun Kwan University, College of Information and Communication
(***) Sung Kyun Kwan University, College of Software

요 약

본 논문에서는 조류 인플루엔자 발생 지역의 기온에 따른 조류 인플루엔자 바이러스 발생의 상관관계를 밝혀내고자 한다. 조류 인플루엔자 바이러스 발생에 영향을 끼치는 외부 환경 요인으로는 기온, 습도, 강수량 등을 들 수 있으나, 본 논문에서는 수집 데이터의 한계로 인해 기온과의 관계만 분석한다. 본 논문에서는 조류 인플루엔자 발생 지역과 해당 지역의 기온 데이터의 연관성을 수치화하는 회귀분석 기법을 이용해 조류 인플루엔자와 외부 기온의 상관관계를 파악, 이를 통해 (1) 조류 인플루엔자 바이러스와 기온의 상관관계 측정, (2) 측정을 통해 향후 연구 혹은 활용방안으로써 활용할 수 있을 것이다.

1. 서론

1.1 연구배경

2016년 11월부터 발생하기 시작한 조류 인플루엔자는 대한민국 사회에 막대한 영향을 끼쳤다. 이전부터 조류 인플루엔자로 인한 피해는 줄곧 있어왔으나 일반 소비자 식탁에까지 그 여파가 미친 건 이번이 처음이다. 계란 값은 며칠 만에 두 배 이상으로 치솟았으며 계란을 이용한 음식이 많은 한국인의 밥상 특성상 그 영향은 이루 말할 수 없을 정도로 막대했다. 닭은 조류 인플루엔자로 인해 대량으로 살처분되었고, 계란은 팔리지 않으니 사회 전반적으로 큰 손실이 있었다. 이를 방지하고자 조류 인플루엔자와 기온과의 상관관계를 분석하여 향후 조류 인플루엔자의 확산을 저지하는 데에 활용할 수 있을 것이다.

1.2 연구 범위 및 수행 절차

본 논문에서는 2014년 4월부터 2017년 6월까지의 조류 인플루엔자 발생 신고 및 기온 데이터를 이용하여 연구를 진행한다. 모든 데이터는 일 단위의 신고 날짜 별로 수집, 분석하였으며 농림축산검역본부 사이트와 기상자료개방포털 사이트에서 데이터를 얻을 수 있었다.

2장에서는 논문 주제와 관련된 연구를 소개하고, 3장에서는 논문 내용 작성을 위해 사용하여야 할 데이터 수집 및 처리 과정을 서술한다. 4장에서는 회귀분석 기법을 통해 조류 인플루엔자 발생 지역과 기온 간의 연관성을 분석하고, 5장에서는 결론을 맺는다.

2. 관련 연구

본 논문의 관련 연구로는 회귀분석 기법이 있다. 회귀분석이란 관찰된 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 기법이다. 영향을 주는 요소인 독립변수와 영향을 받는 요소인 종속변수 간의 관계를 계산하고 이를 통해 multiple R과 같은, 상관관계를 수치로 나타내는 데이터를 도출하여 연관성을 얻을 수 있다. 회귀분석이 사용된 예시로는 날씨와 대중교통 사용량과의 상관관계 또는 최고기온과 빙과류 판매 사이의 상관관계를 분석 등이 있다. 본 논문에서는 기온과 조류 인플루엔자 발생과의 상관관계를 분석하기 위해 회귀분석 기법을 사용한다.

3. 데이터 수집 및 처리

본 논문의 수행 절차는 크게 데이터의 수집 및 처리와 회귀분석 처리 두 단계로 나눌 수 있다. 수집된 데이터 파일에는 본 연구와 관련 없는 데이터가 존재하기 때문에, 본 연구의 첫 번째 단계로는

수집된 데이터 파일로부터 본 연구의 주제와 부합하는 데이터만을 처리한 데이터 파일을 생성하는 단계이고 본 장에서 서술한다. 본 연구의 다음 단계로는 데이터 파일을 회귀분석 기법을 이용하여 조류 인플루엔자 바이러스의 발생과 기온 간의 연관성을 알아내는 단계이며, 이에 대해서는 다음 장에서 서술한다.

3.1 데이터 수집

본 논문에서는 조류 인플루엔자 발생 지역 데이터 파일과 기온 데이터 파일을 필요로 한다. 해당 데이터 파일을 생성하기 위한 관련 데이터는 각각 농림축산검역본부 [3]와 기상자료개방포털 [4]에서 수집하였으며, 이를 통해 조류 인플루엔자 발생 지역 데이터와 해당 지역의 조류 인플루엔자 발생 기간 기온 데이터를 생성하였다. 모든 데이터는 2014년 4월부터 2017년 6월 23일까지 총 3년 2개월간의 일 단위를 기준으로 생성되었다.

3.2 발생 신고 지역 데이터 처리

농림축산검역본부에서 제공하는 데이터 파일로부터 2013년부터 2017년 6월까지의 4년 6개월간의 조류 인플루엔자 발생 신고 데이터를 수집하였으나, 해당 기간에 수집된 데이터 수의 부족 등의 한계로 유의미한 데이터 분석이 불가능한 기간을 제외한 2014년 4월부터 2017년 6월까지의 일 단위 데이터를 이용하였다.

먼저, 수집된 조류 인플루엔자 신고 데이터를 2014년 4월을 시작으로 일 단위로 하나의 데이터 파일에 정리하였다. 하지만 발생 지역이 너무 파편화되어 있어서 인접한 지역을 하나의 큰 구역으로 통합시킬 필요성이 존재하기 때문에 본 논문에서는 조류 인플루엔자 발생 신고 지역을 1~18, 그리고 20의 숫자로 그 구역을 나눴다. 세부사항은 아래 <표 1>과 같다.

<표 1> 조류 인플루엔자 발생 신고 지역 구획 인덱스

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

<Table 1> Compartment index of Avian flu occurrence reported area

각 지역마다 구역을 지정한 이후, 기존의 데이터 파일에서 주소 데이터를 구역의 숫자로 대신함으로써 조류 인플루엔자 발생 신고 지역 데이터 처리하였다.

2.3 기온 데이터 처리

기온 데이터의 경우, 기상자료개방포털에서 제공하는 기온 데이터 파일로부터 3.2절에서 처리된 발생 신고 지역 데이터 파일을 기반으로 이와 상응하는 지역의 기온 데이터를 우선적으로 처리하였다.

그 다음으로는 발생 신고 날짜를 반영하여 당시의 전국 평균 기온과 전국 평균 기온 그리고 발생 지역 기온의 차를 일 단위로 처리하여 최종 데이터 파일을 완성하였다. <표 2>는 최종 데이터 파일의 예시를 보이며, 각 열은 구역(area), 해당 지역의 기온(Temperature), 전국 평균 기온(AvgTemp) 그리고 발생 지역의 기온 차(TempDifferential)로 구성되어 있다.

<표 2> 최종 기온 데이터 예시

area	temperature	AvgTemp	TempDifferential
1	-0.48	0.5	-0.98
5	1.01	0.5	0.51
6	5.56	0.5	5.06
4	4.46	0.5	3.96
4	7.83	0.5	7.33
4	5.75	0.5	5.25
7	-3.44	0.5	-3.94
3	-0.18	0.5	-0.68
7	-1.66	0.5	-2.16
5	-0.25	0.5	-0.75
7	-2.54	0.5	-3.04
15	-5.55	0.5	-6.05
14	12.01	2.5	9.51
11	4.03	2.5	1.53
5	6.29	2.5	3.79
4	1.13	2.5	-1.37
6	3.6	2.5	1.1
5	4.54	2.5	2.04
2	-1.05	2.5	-3.55
8	0.65	2.5	-1.85
7	2.15	2.5	-0.35

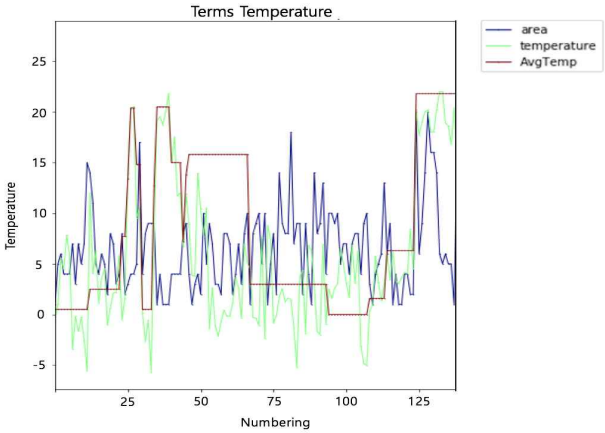
<Table 2> Example of temperature data final vol.

4. 회기분석 처리 및 분석

본 장에서는 앞 장에서 수집 및 처리되어진 두 가지 데이터 파일을 이용하여 상관관계를 분석하였다. 시각적 자료와 통계학적 자료 모두 필요하다고 판단하여 시각적 자료는 Python의 pandas 라이브러리를 이용[5], 통계학적 자료는 회귀분석 기법을 이용해 결론을 도출하고자 하였다. 먼저, 시각적 자료는 (그림 1)과 같다.

(그림 1)에서 area를 의미하는 파란 선은 조류 인플루엔자 발생 구역이고, temperature를 의미하는 연두 선은 해당 지역 당시 기온, AvgTemp를 의미하는 빨간 선은 당시의 전국 평균 기온이다.

(그림 1) Pandas를 통해 작성된 선 그래프



(Figure 1) Line graph drawn with pandas

(그림 1)의 시각적 자료만으로는 조류 인플루엔자 발생 지역과 기온과의 상관관계를 명확하게 얻을 수 없는 한계가 존재한다. 그래서 본 논문에서는 객관적인 수치로 연관성을 얻기 위해 회귀분석 기법을 이용하여 분석하였다[6]. 회귀분석 기법 적용을 위한 독립변수로는 조류 인플루엔자 발생 구역, 종속변수는 전국 평균 기온과 발생 지역 당시 기온의 차로 지정하였다. 결과는 아래 <표 3>과 같다.

<표 3> 회귀분석 기법 적용 결과표

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.4451122
R Square	0.1981249
Adjusted R Square	0.1923142
Standard Error	7.5313122
Observations	140

<Table 3> Result table by regression analysis technique

분석 결과 Multiple R이 0.445인 것으로 드러났다. 이는 기온 차가 조류 인플루엔자 발생에 약 44.5%의 영향을 준다는 뜻이다. 한 요소가 다른 요소에 의해 절반이 가까운 확률로 영향을 받는다면 이 두 가지 요소 사이에는 적지 않은 연관성이 있다는 것을 의미한다. 이를 단편적으로 보여주는 예시로는 암의 발생률과 사망률의 관계가 있다. 대중적으로 어떤 사람에게 암이 발생하면 그 사람이 사망할 가능성이 매우 높다고 생각한다. 하지만 실제 암에 의한 사망률은 27.8%로[7], 즉, 암과 사망률 간의 연관성은 27.8에 불과한 것이다. 이처럼 전체 영향의 절반도 되지 않는 확률로도 두 요소의 상관관계가 높다고 여겨진다. 따라서 본 연구의 분석 과정에서 회귀분석을 통하여 얻은 객관적 수치를 통해 조류 인플루엔자 바이러스와 기온 간의 상관관계가 높음을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 조류 인플루엔자 발생 신고 지역과 기온과의 상관관계를 분석하였다. 발생 신고 지역 데이터와 지역 기온 데이터, 그리고 전국 평균 기온 데이터를 정리하여 데이터 파일에 정리하였고, 회귀분석을 통해 데이터 간의 연관성을 multiple R, 즉 44.5%라는 객관적인 수치로 증명하여 결론을 도출하였다.

향후, 교차검증과 주성분 분석 기법과 같이 더욱 다양한 기법, 그리고 습도, 날씨 등의 다양한 주변 환경 데이터를 기반으로 분석하여 나은 결론을 도출할 수 있을 것으로 기대된다. 이에 분석 기법을 많이 연구해서 다양한 데이터로 본 논문의 결론을 보강할 계획이다.

참고 문헌

[1] 농림축산검역본부, “신고발생현황”, Available: http://www.qia.go.kr/animal/prevent/listwebQiaCom.do?type=2_12qlgzls (2017-09-16 방문)

[2] 기상자료개방포털. “실황분석자료 데이터", Available: https://data.kma.go.kr/data/rmt/rmtList.do?code=400&pgmNo=570 (2017-09-16 방문)

[3] Jaime Sanches, “Data Visualization with Plotly and Pandas”, SODA Developers, Available: https://dev.socrata.com/blog/2016/02/02/plotly-pandas.html, 02 Feb 2016.

[4] Kim Gwangseob, Lee Gichun, “Estimate Extreme Hydrologic Event at Seoul Using Regression Analuses”, Journal of the Korean Society of Hazard Mitigation, VoL. 12, No. 3, pp. 263~270, 2012.

[5] 국가암정보센터, “주요암 사망분율”, Available: http://www.cancer.go.kr/mbs/cancer/subview.jsp?id=cancer_040201000000, 2017년 10월 26일.