

2020 하계 환경연수프로그램

빅데이터를 이용한 한국 사회의 가뭄, 홍수 대응 분석

포항공과대학교 화학공학과 송지암

포항공과대학교 수문기후 연구실

지도 교수: 감종훈 교수님

I. Background

가뭄과 홍수는 우리 삶에 밀접하게 물과 연관되어 있는 자연재해이다. 또한, 가장 쉽게 접할 수 있는 자연재해이기도 하다. 최근에는 장마철 홍수로 인해 홍수에 대한 관심이 높아지고 있는 상황이다. 그동안의 연구들을 보면 이 두 자연재해에 대한 메커니즘과 물리적인 특성들을 파악하는데 집중 되어있다. 그렇기 때문에 사람은 자연스레 이 연구 과정에서 빠지게 되었다. 하지만, 이런 자연재해는 사람들에게도 밀접하기 때문에 큰 영향을 미치게 된다. 따라서, 가뭄, 홍수 두 자연재해와 이에 대응하는 사람들을 동시에 엮어서 분석하고 이해하는 과정이 필요하게 된다. 이는 사람들에게 미치는 영향만을 분석하는 것을 넘어서, 이런 사람들의 반응들이 가뭄, 홍수의 메커니즘에 끼치는 영향을 분석해 자연재해의 메커니즘을 더 정확하게 분석할 수 있게 되고, 이를 토대로 자연재해에 대응하는 정책에 대한 솔루션을 제시하는데 큰 도움을 줄 수 있을 것이다.

II. Progress

사람들의 반응을 살피기 위해 가장 좋은 방법은 인터넷에서 무엇을 검색하는지 알아보는 것이다. 구글에서는 구글을 통해 검색되는 검색어의 동향을 설정된 기간 동안 데이터로 보여주는 구글 트렌드 서비스를 제공하고 있다. 인터넷 시대인 요즘 시대에 구글 트렌드를 통해 가뭄, 홍수에 대한 인지도를 유추할 수 있을 것이다. 네이버에서도 비슷한 서비스로 NAVER Data lab 을 제공하고 있다. 구글의 경우 2004 년부터의 데이터를, 네이버의 경우 2016 년부터의 데이터를 제공하고 있다. 이 두 가지 서비스를 이용해 사람들의 반응을 측정할 것이다.

가뭄, 홍수의 물리적인 측정의 경우 SPI(Standard Precipitation Index)를 이용한다. 강수량의 경우 비가 오지 않는 날이 더 많기 때문에 데이터가 0 으로 치우쳐진 분포를 보이게 된다. 이런 분포를 바로 정규화하기 어려우므로 감마분포 가정을 통해서 정규화를 진행하고 난 후 Index 를 얻은 것이 SPI 이다. SPI 의 경우 강수량을 측정하는 특정 시간 단위에 따라 SPI 3,6,9,12 등으로 나눈다. 짧을수록 단기간, 길수록 장기간의 강수 동향을 살필 수 있다.

이번 분석의 경우 한국사회에 맞는 분석을 진행할 것이다. 구글트렌드의 경우 한국에서 인터넷, 핸드폰의 보급으로 인터넷 사용량이 어느정도 늘어난 2009 년부터 유의미한 값을 가지므로 2009 년부터 분석을 진행한다. 네이버의 경우 2016 년부터

데이터를 제공하므로 그에 맞는 데이터부터 분석을 진행하고 구글트렌드와 비교할 것이다. SPI 의 경우 1904 년부터 2018 년 기간의 SPI3, SPI 12 데이터를 이용할 것이다.

III. Result

A. Social response

구글 트렌드, NAVER Data lab 의 경우 Python 을 이용해서 데이터를 쉽게 불러오고 plot 하고 저장할 수 있도록 만들어 사용했다. 구글 트렌드는 pytrend package 를 사용했고, NAVER Data lab 은 제공하는 API 를 통해 데이터를 가져왔다. R 에서도 gtrendR 과 같은 구글 트렌드 API 관련 package 가 존재했으나 제공하는 기능이 적었다. 구글 트렌드의 유의미한 분석기간인 09 년도부터 20 년도 까지 가뭄, 홍수에 대한 검색어 동향은 다음과 같다.

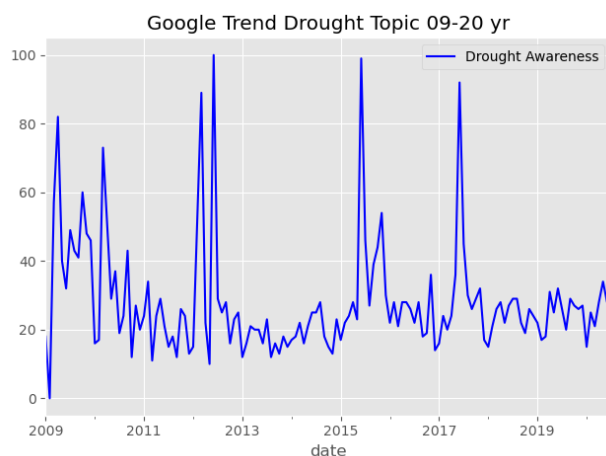


FIGURE 1. DROUGHT AWARENESS 09-20

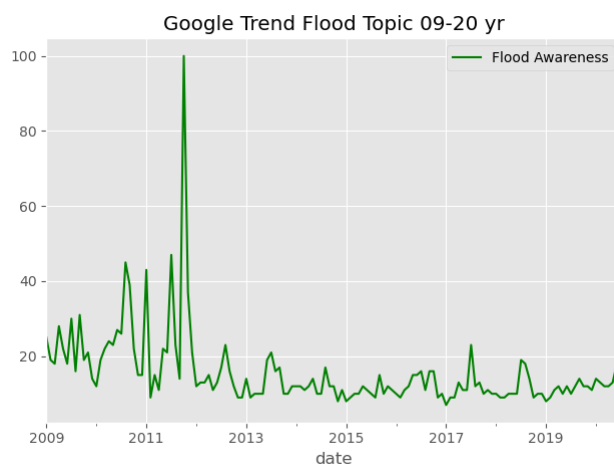


FIGURE 2. FLOOD AWARENESS 09-20

가뭄의 경우 12,15,17 년도에 유의미한 peak 가 발견되었다. 09-11 년도의 data 의 경우 한국의 경우 초기데이터에 가깝기 때문에 데이터 불안정성이 있을 수 있으므로 제외한다. 위 3개의 peak 를 통해 물리적으로도 저 시기에 가뭄이 있었음을 확인해 비교할 수 있을 것이다.

홍수의 경우 11 년도에 큰 peak 가 나타나고 있다. 이는 11 년도에 일어난 태국 홍수로 인해 인지도가 증가한 것으로 보인다. 당시 태국 홍수의 규모는 상당히 커서 세계적으로 이목을 끌었고 그로 인한 peak 로 보인다. 이로 인해 다른 peak 가 나타나지 않으므로 이를 제외하고 가뭄에서도 peak 가 나타나는 12 년도부터 비교를 통해 유의미한 분석을 진행을 해보았다.

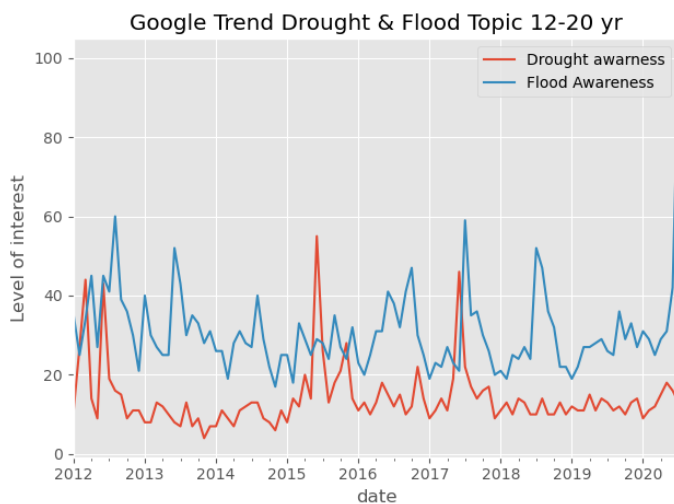


FIGURE 3. DROUGHT & FLOOD TOPIC 12-20

두 재해를 동시에 비교해보았을 때, 평균적으로 홍수에 대한 인지도가 가뭄에 대한 인지도보다 높은 것을 알 수 있었다. 또한, 홍수의 경우 주기적으로 여름철에 인지도가 높아지는 것을 관찰 할 수 있었다. 살펴볼 것은 두 재해에 대해 동시에 인지도변화를 살펴보는 것이다. 12, 17 년도의 경우 가뭄에 대한 peak 가 나타난 후 홍수에 대한 peak 가 나타났다. 반면, 15 년도의 경우에는 가뭄에 해당하는 peak 만 나타났다. 따라서, 이 두 가지 pattern 에 대한 추가 분석을 통해 차이점을 알아보기로 했다. 현재 세계적으로 가뭄 뒤에 홍수가 오는 extreme to extreme 이 증가하고 있다.¹ 따라서, 인지도로부터 나타나는 extreme to extreme 패턴이 실제로 일어났는지 SPI 를 통해 알아볼 수 있을

¹ He, X., & Sheffield, J. (2020). Lagged compound occurrence of droughts and pluvials globally over the past seven decades. *Geophysical Research Letters*, 47, e2020GL087924.

것이다. 또한, 사회의 반응을 추가적으로 분석해 해당 패턴에 대한 사람들의 반응에 대해서도 알아볼 수 있을 것이다.

B. SPI3 (Find drought & flood)

물리적인 가뭄, 홍수 발생을 찾기 위해 SPI 데이터틀 이용한다. 현재 분석에 사용할 기간이 12-20 년도이므로 장기적인 가뭄 detect 에 유용한 SPI12 보다는 단기 가뭄과 짧은 기간 동안 일어나는 홍수에 맞는 SPI3 를 이용한다. 가뭄의 경우 SPI3 기준으로 -0.8 부터 onset, 이후 +0.2 를 넘어가면서 recovery 기간을 거친다. 이를 통해 가뭄을 찾을 수 있을 것이다. 홍수의 경우 반대로 0.8 을 onset, -0.2 를 recovery 로 두고 찾는다.

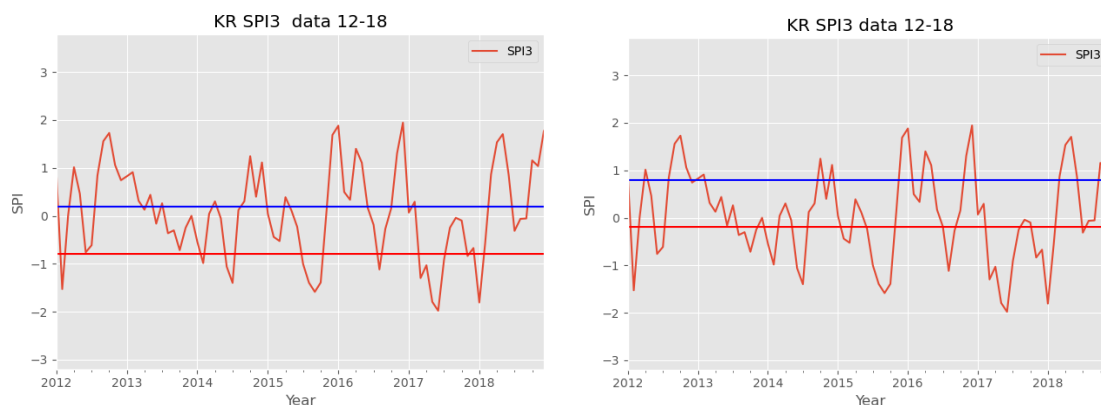


FIGURE 4. SPI3 DATA 12-18

가뭄의 경우 12, 14(2 번), 15, 16, 17, 17~18 년도로 6 번 관찰 되었다. 반대로 홍수의 경우 12(2 번), 14, 15~16, 16~17, 18 년도로 5 번 관찰 되었다. 12 년도의 경우 구글 트렌드 데이터에서 보았듯이 가뭄, 홍수가 모두 관찰 되었고 가뭄 뒤 홍수가 나타나는 것이 일치한다. 15 년도의 경우 가뭄만 나타나는 것으로 또한 일치했다. 17 년도의 경우 SPI3 데이터로부터는 홍수는 일어나지 않았다. 그럼에도 Flood awareness 가 증가하였다. 이 경우 홍수를 더 잡아내기 쉬운 SPI1 데이터를 통해 추가 비교를 진행이 필요할 것으로 보인다.

C. NAVER Data lab & Google trend

NAVER 의 경우 google 에 비해 짧은 기간의 데이터를 제공하고 있지만 현 연구주제는 한국에 집중하고 있기 때문에 질적인 측면에서는 google 에 앞설 것이다. 따라서, google trend 의 data 들을 어느정도 NAVER Data lab 과 비교하고 이를 통해 보정을 할 필요성이 존재한다. 보정을 하기 위해서는 어느정도 일치하는 면이 존재해야한다. 따라서, 두 data 를 비교해본다.

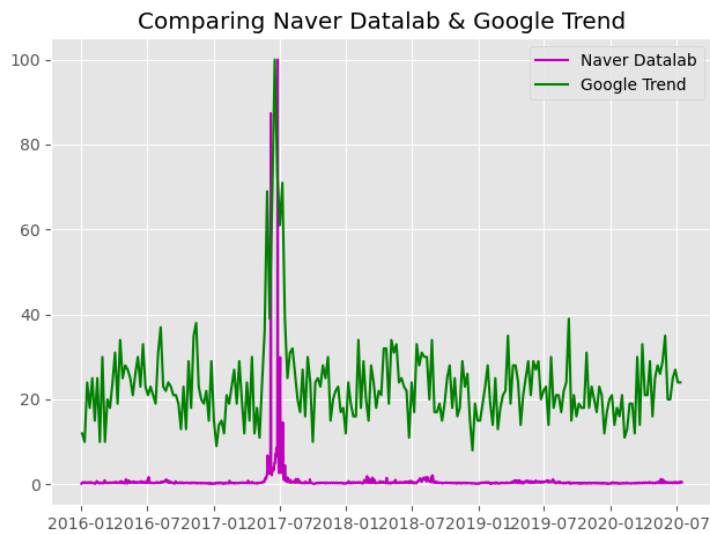


FIGURE 5. COMPARING WITH GOOGLE TREND AND NAVER DATALAB

NAVER 에서 16 년도부터 데이터를 제공하므로 16 년도부터 20 년도까지의 데이터를 구글 트렌드와 비교했다. NAVER 의 경우 date scale, 구글 트렌드의 경우 두 데이터 모두 17 년도 6 월 부근에서 peak 를 보이는 것을 확인할 수 있었다. 두 데이터의 큰 차이로는 peak 이외의 부분이다. NAVER 의 경우 0 에 가까운 값들을 보이는 반면, 구글 트렌드는 평균 20 정도의 값을 유지한다. 그럼에도 peak 를 보이는 분포 형태 자체는 비슷하다. 때문에 두 분포를 서로 보정해 이용하는 것에는 큰 영향을 미치지 않을 것이다. 이를 이용해 구글 트렌드 값들을 보정해 한국에 맞는 값들로 분석을 이어나갈 수 있을 것이다.

IV. Conclusion

구글 트렌드와 SPI 데이터를 통해 실제 일어난 가뭄, 홍수와 그에 대한 사람들의 반응을 묶어서 살펴 볼 수 있었다. 구글 트렌드 비교가 유의미한 구간인 12-20 년도에 맞춰 분석을 진행했고, SPI는 SPI3 데이터를 이용해 가뭄, 홍수를 detection 했다. 그 결과, 12, 15, 17 년도에 가뭄 인지도에 유의미한 peak 를 관찰 할 수 있었고, 홍수의 경우 여름철에 주기적으로 인지도가 높아지는 것을 볼 수 있었다. 그 중에서 12, 17 년도는 가뭄의 인지도가 높아진 직후, 홍수의 인지도가 높아지는 것을 관찰 할 수 있었고, 15 년도의 경우 가뭄의 인지도만 증가한 것을 확인 할 수 있었다. 이를 실제 SPI3 데이터와 비교 했을 때, 12 년도의 경우에는 실제로도 가뭄 이후 홍수가 일어난 것을 확인 할 수 있었다. 15 년도의 경우에도 가뭄만 일어난 것을 확인 할 수 있었다. 반면, 17 년도의 경우 실제 홍수는 일어나지 않았음을 확인할 수 있었다. 이를 추가적으로 분석하기 위해서는 홍수를 잘 detect 하는 SPI1 데이터를 통해 추가적으로 홍수에 대한

분석을 해보고, 이 후 결과를 바탕으로 분석을 진행할 것이다. 또한, 추후 분석을 위해 구글 트렌드와 NAVER datalab 데이터를 비교해보았고, 서로 유사한 분포를 보이는 것을 바탕으로 보정을 통한 값으로 NAVER datalab 의 데이터가 없는 값들을 구글트렌드를 통해 대체 가능하다는 것을 알 수 있었다. 이는 추후 한국에 최적화된 분석에 사용할 수 있을 것이다. 또한, NAVER datalab 에서 제공하는 추가적인 trend 분석을 구글 트렌드와 비교해 보정 후 분석을 해볼 수 있을 것이다.