



기상청

2021 날씨 빅데이터 콘테스트

기상에 따른 산림재해 예측

김상완, 최하영



CONTENTS

I. 분석 개요

II. 데이터 로딩

III.. 데이터 탐색

IV. 데이터 처리

V. 모형 구축

VI. 모형 검증

분석 개요

연구 배경

01

기후변화로 인한 강수량 및 강우 패턴 변화

- 집중호우일수 증가와 강한 태풍 비중이 높아질 것 (기상청, 2013)
- 여름철 호우 재해의 발생 빈도 연평균 5.3회 → 8.8회 증가 (국립기상연구원, 2007)

02

산림 재해로 인한 경제적 피해

- 태풍·호우로 인한 재산 피해액은 1조 2,585억 원으로 최근 10년 연평균 피해액의 약 3배에 달함 (기상청, 2020)
- 6,175건의 산사태 발생으로 역대 3번째였던 것으로 나타남 (기상청, 2020)

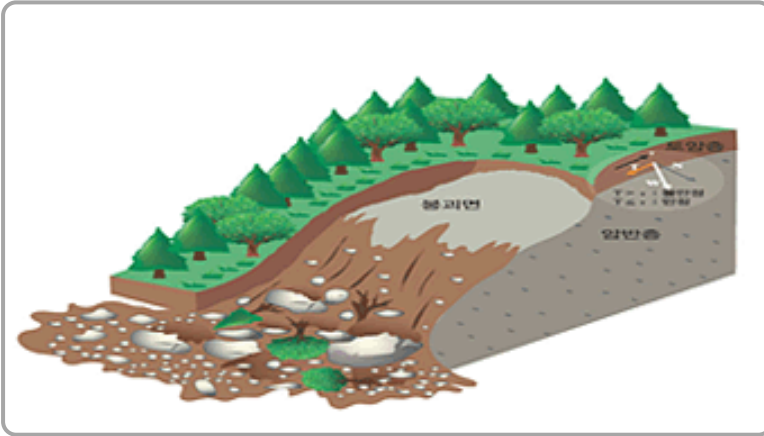
03

산사태 예측 기술 개발 · 개선

- 산사태 발생 시 토석류 피해 범위 정밀하게 예측 가능한 수치 모델 개발 (한국농어촌공사, 2021)
- 산사태 예측 조기경보시스템 개발 (한국지질자원연구원, 2020)

분석 개요

산사태 정의 (산림청, 2019)



[산사태]



[토석류]

- 산사태 : 자연적 또는 인위적인 원인으로 산지가 일시에 붕괴되는 것
- 토석류 : 산지 또는 계곡에서 토석 · 나무 등이 물과 섞여 빠른 속도로 유출되는 것
- 빗물이 산지사면의 토양내부로 침투하여 포화도가 증가함에 따라 불투수층(암반)과 흙의 경계가 분리되어 토층(암반상층)의 흙이 떨어져 나가는 현상

분석 절차

분석 목표 및 분석 계획

목표

- 2020년 6월 ~ 2020년 9월 경상도 지역 산사태 분석 및 예측을 통해 산사태 발생 예측 모델 구축
- 현재의 산사태 발생가능성 모형을 바탕으로 의미있는 변수를 도출하고, 미래에 대한 발생가능성 분석 및 산사태 예측 기술 개발 및 개선



활용 데이터 상세 내용

산사태 예측 모형 구축에 사용된 데이터 정보

임상도

- 임향을 개략적으로 알아볼 수 있도록 작성된 도면
- 임종, 임상, 경급, 영급, 소밀도 등을 구분하여 편집

토양도

- 토양을 분류해 이것을 지도상에 나타낸 것
- 토양의 성분 및 성질을 밝혀 그 분포 상태를 나타낸 지도

임도

- 임산물의 운반 및 산림의 경영관리상 필요로 설치한 도로



기상

- 일강우량: 하루 24시간동안 집계된 강우량
- 최대강우강도: 단위 시간당 강우량을 측정한 것
- 최대순간풍속: 단위 시간당 순간 풍속의 최댓값

행정동

- 행정편의를 위하여 설정한 행정구역 단위
- 행정동코드(10) = 시도(2) + 시군구(3) + 읍면동(5)

법정동

- 법률로 지정된 행정구역 단위
- 법정동코드(10) = 시도(2) + 시군구(3) + 읍면동(3) + 리(2)

제공 데이터

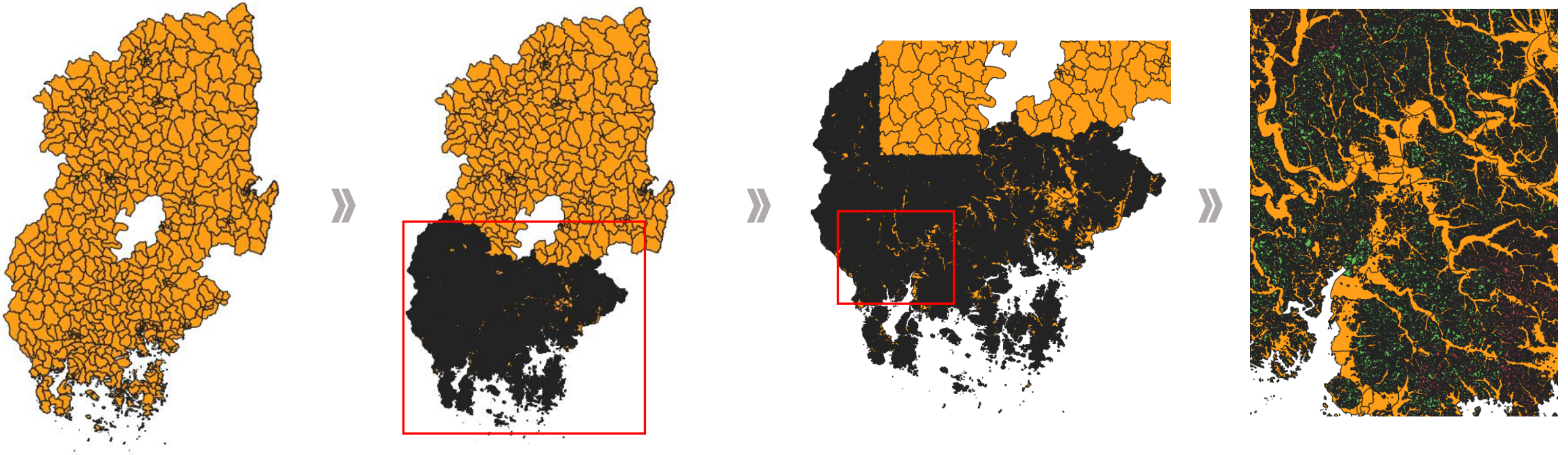
산사태 발생 이력에 주어진 변수 정보

date	sd	sgg	umd	sum_cnt	sum_hpa	행정동코드	법정동코드
20110709	경상남도	밀양시	내일동	1	1.2	4827051000	4827010100
20110709	경상남도	밀양시	단장면	4	3.7	4827035000	...
20110709	경상남도	밀양시	무안면	5	4.2	4827038000	...
20110709	경상남도	밀양시	부북면	6	7.8	4827031000	...
20110709	경상남도	밀양시	산외면	1	2	4827032000	...
...

- 제공받은 '산사태 발생 이력.csv' 파일의 행정구역 경계는 행정동으로 제공
- 산림공간정보와 매칭시키기 위해 행정동코드를 구한 뒤, 법정동코드로 merge
- 산림 공간 정보와 산사태 발생 이력 데이터를 merge시키기 위해 행정동코드와 법정동코드를 1대1 매핑
- (출처: 국가공간정보포털)

데이터 전처리

산림 공간 정보 전처리 과정



- 법정동 행정구역경계에 임상도, 토양도, 임도를 찍어 겹치는 곳을 추출
- 임상도의 값이 존재하는 곳을 산림 지역으로 선택한 뒤, 임도에 해당하는 부분 제외
- GIS 툴인 ArcGIS 이용하여 산림 공간 정보 속성 추출

데이터 이해

시간축이 다른 데이터를 병합시키기 위한 과정

산사태 발생 이력

- 과거에 발생한 산사태 이력으로 2011년 - 2019년 사이 발생한 이력
- 이때의 임상 및 토양의 상태는 알 수 없음

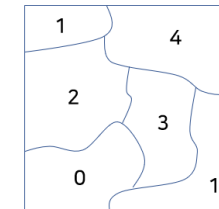
임상도 및 토양도 (1:5,000)

- 제공된 임상도 및 토양도가 나타내는 속성은 현재 시점
- 시점이 다른 데이터를 병합시켜 의미있는 변수를 만들고자 함
- (출처: FGIS 산림공간정보서비스)

해결 방법

- 주어진 point reference data를 areal data로 변형
- 지점이 아닌 지역에 대해 전체 면적 대비 각 산림 특성 값의 면적 비율(%)로 가공

<예시: 임상도 임상코드(FRTP_CD)>



A 지역 임상코드(FRTP_CD) 특성 수치화

FRTP_CD_1 = 15.0306
FRTP_CD_2 = 64.6136
FRTP_CD_3 = 16.3658
FRTP_CD_4 = 0.0602
FRTP_CD_0 = 3.9297

- 특정 지점이 아닌 행정동 내 토양 및 임상 평균적 특성을 수치화
- 전체 면적 대비 각 산림 특성값의 면적 비율(%)을 산출

데이터 전처리

기상정보데이터 설명

	A	B	C	D	E	F	G
1	지점번호 ▾	경도 ▾	위도 ▾	노장해발고도 ▾	지점명 ▾	예보구역코드 ▾	법정동코드 ▾
37	155	128.5729	35.1702	37.15	창원	11H20301	4812552000
41	162	128.4356	34.84546	32.67	통영	11H20401	4822010900
57	192	128.04	35.16379	30.21	진주	11H20701	4817012600
79	253	128.893	35.22666	59.34	김해시	11H20300	4825010300
81	255	128.6275	35.2264	46.77	북창원	11H20301	4811012100
82	257	129.02	35.3072	14.85	양산시	11H20102	4833031027
88	263	128.2881	35.3226	14.18	의령군	11H20600	4872025024
89	264	127.7454	35.5114	151.2	함양군	11H20502	4887025022
101	284	127.909	35.6674	225.95	거창	11H20502	4888025028

- 빅데이터 분석 플랫폼 날씨마루에서 HIVE 이용하여 방재기상관측(AWS)지점 추출
- 방재기상관측(AWS)지점은 법정동코드를 제공
- 경상북도, 경상남도에 위치하는 AWS 지점만 추출

데이터 전처리

기상정보데이터 설명

파일명	파일 설명	변수명	변수 설명	형식	예제
AWS_HR_RN	DB:AWS: 통계:강수	TM	관측시각	STRING	1999-01-01 00:00:00.0
		AWS_ID	AWS 번호	BIGINT	96
		RN_DAY	일강수량	DOUBLE	5.6
		RN_HR1	1시간 강수량	DOUBLE	0.2
AWS_HR_WD	DB:AWS: 통계:바람	TM	관측시각	STRING	1999-01-01 00:00:00.0
		AWS_ID	AWS 번호	BIGINT	96
		WS_INS_MAX	최대순간풍속	DOUBLE	4.5
DB_AWS_RN_TIM	DB:AWS: 강수:시	TMA	관측시각	STRING	1997-01-01 00:00:00.0
		STN_ID	지점번호	BIGINT	151
		HR1_MAX_RN	1시간 최대 강수량	DOUBLE	7.6

- 빅데이터 분석 플랫폼 날씨마루에서 HIVE를 통해 경상도에 위치하는 방재기상관측(AWS)지점의 값들을 추출
- AWS_HR_RN.RN_DAY 와 AWS_HR_RN.RN_HR1 를 이용하여 강수량 산출
- AWS_HR_WD.WS_INS_MAX를 이용하여 일 최대순간풍속,
- DB_AWS_RN_TIM.HR1_MAX_RN을 이용하여 일 최대강우강도 산출

데이터 전처리

기상 정보 전처리 과정

	지점번호	법정동코드	umd	sgg	sd
1	90	4282033035	42820330	42820	42
2	92	4283032021	42830320	42830	42
3	95	4278025624	42780256	42780	42
4	96	4794025027	47940250	47940	47
5	98	4125010300	41250103	41250	41
6	99	4148025025	41480250	41480	41
7	100	4276037025	42760370	42760	42
8	101	4211011800	42110118	42110	42
9	102	2872033024	28720330	28720	28
10	104	4215036027	42150360	42150	42
11	105	4215010700	42150107	42150	42



	key	지점번호
99	경상남도 밀양시 청도면	922
100	경상남도 밀양시 초동면	927
101	경상남도 밀양시 초동면	288
102	경상남도 밀양시 초동면	922
103	경상남도 사천시 곤명면	907
104	경상남도 사천시 곤명면	917
105	경상남도 사천시 곤양면	907
106	경상남도 사천시 곤양면	917
107	경상남도 사천시 서포면	907
108	경상남도 사천시 서포면	917
109	경상남도 사천시 용현면	917

- 방재기상관측(AWS)지점은 법정동코드로 제공
- 법정동코드, umd, sgg, sd 순으로 방재기상관측(AWS)지점의 지점번호를 매핑시킴
- 하나의 key 값에 대해 여러 지점번호가 할당될 경우, 추후에 지점번호에 해당하는 기상 정보의 평균을 할당

데이터 전처리

기상정보 전처리 과정

	key	date	aws_id	cs0	cs1	cs2	cs3
1	경상남도거창군가북면	2011-08-09	284	515.0	1044.0	118.5	0.0
2	경상남도거창군가북면	2011-08-09	946	435.5	0.0	0.5	7.0
3	경상남도거창군가북면	2011-11-30	284	273.5	0.0	0.0	0.0
4	경상남도거창군가북면	2011-11-30	946	341.0	0.0	0.0	0.0
5	경상남도거창군가북면	2012-04-25	284	288.5	0.0	1.0	70.5
6	경상남도거창군가북면	2012-04-25	946	390.0	0.0	1.5	89.0
7	경상남도거창군가북면	2012-06-12	284	0.0	0.0	0.0	15.0
8	경상남도거창군가북면	2012-06-12	946	8.0	0.0	0.0	26.0
9	경상남도거창군가북면	2012-07-06	284	317.0	277.0	0.0	0.0
10	경상남도거창군가북면	2012-07-06	946	542.5	555.5	0.0	0.0



	key	date	aws_id	cs0	c1	c2	c3
1	경상남도거창군가북면	2011-08-09	284	515.0	1559.0	1677.5	1677.5
2	경상남도거창군가북면	2011-08-09	946	435.5	435.5	436.0	443.0
3	경상남도거창군가북면	2011-11-30	284	273.5	273.5	273.5	273.5
4	경상남도거창군가북면	2011-11-30	946	341.0	341.0	341.0	341.0
5	경상남도거창군가북면	2012-04-25	284	288.5	288.5	289.5	360.0
6	경상남도거창군가북면	2012-04-25	946	390.0	390.0	391.5	480.5
7	경상남도거창군가북면	2012-06-12	284	0.0	0.0	0.0	15.0
8	경상남도거창군가북면	2012-06-12	946	8.0	8.0	8.0	34.0
9	경상남도거창군가북면	2012-07-06	284	317.0	594.0	594.0	594.0
10	경상남도거창군가북면	2012-07-06	946	542.5	1098.0	1098.0	1098.0

- 각 key값에 대하여 지점번호와 날짜에 해당하는 일강수량 추출한 후, 일 누적 강수량 산출
- cs_i: i일 전 일강수량 → c_i: i일 누적 강수량

데이터 전처리

기상 정보 전처리 과정

	key	date	aws_id	cs0	c1	c2	c3	max_rn	max_hr1	max_wd
1	경상남도거창군가북면	2011-08-09	284	515.0	1559.0	1677.5	1677.5	16.5	16.0	6.3
2	경상남도거창군가북면	2011-08-09	946	435.5	435.5	436.0	443.0	24.5	18.0	8.6
3	경상남도거창군가북면	2011-11-30	284	273.5	273.5	273.5	273.5	5.5	4.5	7.6
4	경상남도거창군가북면	2011-11-30	946	341.0	341.0	341.0	341.0	9.5	9.0	6.8
5	경상남도거창군가북면	2012-04-25	284	288.5	288.5	289.5	360.0	8.5	7.0	8.6
6	경상남도거창군가북면	2012-04-25	946	390.0	390.0	391.5	480.5	9.0	9.0	8.6
7	경상남도거창군가북면	2012-06-12	284	0.0	0.0	0.0	15.0	0.0	0.0	8.0
8	경상남도거창군가북면	2012-06-12	946	8.0	8.0	8.0	34.0	8.0	8.0	11.2
9	경상남도거창군가북면	2012-07-06	284	317.0	594.0	594.0	594.0	9.0	8.0	10.7
10	경상남도거창군가북면	2012-07-06	946	542.5	1098.0	1098.0	1098.0	15.5	14.5	7.7
11	경상남도거창군가북면	2012-08-12	284	204.0	218.0	315.0	315.0	27.0	27.0	18.9

- 각 key에 대하여 지점번호와 날짜에 해당하는 기상 정보 merge
- c_i : i 일 전 누적강수량, max_rn : 일 최대강우강도, max_hr1 : 일 최대 1시간 강수량, max_wd : 일 최대순간풍속
- 하나의 key 값에 대해 여러 지점번호가 할당된 경우, 지점번호에 해당하는 기상 정보의 평균을 할당

데이터 전처리

기상정보 전처리 과정

	new_key	rn0	rn1	rn2	rn3	rn_tim	rn_hr1	wd
1	경상남도거창군가북면2011-08-09	475.25	997.25	1056.75	1060.25	20.50	17.00	7.45
2	경상남도거창군가북면2011-11-30	307.25	307.25	307.25	307.25	7.50	6.75	7.20
3	경상남도거창군가북면2012-04-25	339.25	339.25	340.50	420.25	8.75	8.00	8.60
4	경상남도거창군가북면2012-06-12	4.00	4.00	4.00	24.50	4.00	4.00	9.60
5	경상남도거창군가북면2012-07-06	429.75	846.00	846.00	846.00	12.25	11.25	9.20
6	경상남도거창군가북면2012-08-12	121.50	135.00	216.75	216.75	16.00	16.00	13.70
7	경상남도거창군가북면2012-08-13	1575.50	1697.00	1710.50	1792.25	30.25	27.75	6.45
8	경상남도거창군가북면2012-09-05	296.50	371.75	371.75	376.25	12.50	11.50	7.10
9	경상남도거창군가북면2012-09-16	1085.25	1086.00	1096.25	1125.00	13.50	12.00	10.10
10	경상남도거창군가북면2012-09-17	1468.25	2553.50	2554.25	2564.50	31.50	24.75	17.75
11	경상남도거창군가북면2018-08-24	1058.85	1652.20	1652.20	1652.20	17.50	16.15	12.75
12	경상남도거창군가북면2018-12-04	162.30	294.10	294.10	294.10	6.10	5.30	10.10
13	경상남도거창군가북면2019-03-20	29.90	29.90	29.90	29.90	4.60	4.35	10.25

- 날짜와 지역을 구분하기 위한 새로운 키 생성 : new_key := (key, date)
- new_key를 기준으로 그룹화하여 기상정보의 평균을 계산하여 할당

데이터 전처리

generator

0/1 generator

- 산사태 발생 이력만 주어진 상황에서 미발생 이력을 임의로 생성
- 산사태 발생/미발생을 1:3의 비율로 생성

date generator

- 2010년 - 현재까지의 매일 기상 데이터 추출
- 그 중, 제공받은 산사태 발생 이력의 일자를 제외한 날들은 산사태가 발생하지 않았다고 가정
- 산사태가 발생하지 않은 일자들 중, **시간당 강수량이 2mm 이상 & 최대순간풍속이 7m/s 이상**인 날들을 1순위로 일자 추출
- 1순위 후보군이 부족한 경우에는 시간당 강수량이 2mm 이상인 날들(2순위) 중 일자 추출

데이터 전처리

기상 정보에 대한 일정기준

시간당 강수량
2mm

- 바람이 불지 않는다 가정했을 때, 톨 사이즈(355ml) 잔으로 약 7잔 정도의 비를 받을 수 있는 정도
- 우산을 쓰지 않는 경우, 옷이 많이 젖을 수준
- (출처: 기상청)

최대순간풍속
7m/s

- 풍력계급 : 3, 명칭 : 산들바람
- (육상상태) 나뭇잎과 가는 가지가 쉴 새 없이 흔들리고 깃발이 가볍게 휘날리는 수준
- (해상상태) 물결끝이 부서지며 거품이 생기고, 여기저기 흰 파도가 나타나는 수준
- (출처: 보퍼트 풍력계급)

데이터 전처리

date generator

후보군

- 1순위 후보군 : 시간당 강수량이 2mm 이상 & 최대순간풍속이 7m/s 이상인 날
- 2순위 후보군 : 시간당 강수량이 2mm 이상인 날

이전 산사태 발생 이력 없음

- $\text{sum_cnt} < 6$: 후보군 중 31개 중에서 $\text{sum_cnt} \times 3$ 만큼 추출
- $\text{sum_cnt} \geq 6$: 후보군 중 183개 중에서 $\text{sum_cnt} \times 3$ 만큼 추출
- 비교적 최근 일자에서 뽑기 위함

이전 산사태 발생 이력 있음

- $\text{sum_cnt} \times 3 > \text{diff}$: 복원추출을 통해, 이전 산사태 발생 일자와의 차이에서 추출
- $\text{diff} \geq \text{sum_cnt} \times 3 > \#candi$: 1순위 후보군에서 최대한 뽑고, 2순위 후보군에서 마저 추출
- $\text{sum_cnt} \times 3 \geq \#candi$: 산사태 발생 이력 없는 경우와 동일

데이터 전처리

date	sd	sgg	umd	key	sum_cnt
2011-07-09	경상남도	밀양시	청도면	경상남도밀양시청도면	2
2011-07-09	경상남도	밀양시	초동면	경상남도밀양시초동면	1
...



date	sd	sgg	umd	key	sum_cnt
2011-07-09	경상남도	밀양시	청도면	경상남도밀양시청도면	1
2011-07-09	경상남도	밀양시	청도면	경상남도밀양시청도면	1
2011-07-09	경상남도	밀양시	초동면	경상남도밀양시초동면	1
...



date	sd	sgg	umd	key	event
2011-07-09	경상남도	밀양시	청도면	경상남도밀양시청도면	1
2011-06-25	경상남도	밀양시	청도면	경상남도밀양시청도면	0
2011-06-22	경상남도	밀양시	청도면	경상남도밀양시청도면	0
2011-05-10	경상남도	밀양시	청도면	경상남도밀양시청도면	0
2011-07-09	경상남도	밀양시	초동면	경상남도밀양시초동면	1
2011-07-07	경상남도	밀양시	초동면	경상남도밀양시초동면	0
2011-07-04	경상남도	밀양시	초동면	경상남도밀양시초동면	0
2011-04-22	경상남도	밀양시	초동면	경상남도밀양시초동면	0
...

분석 시나리오

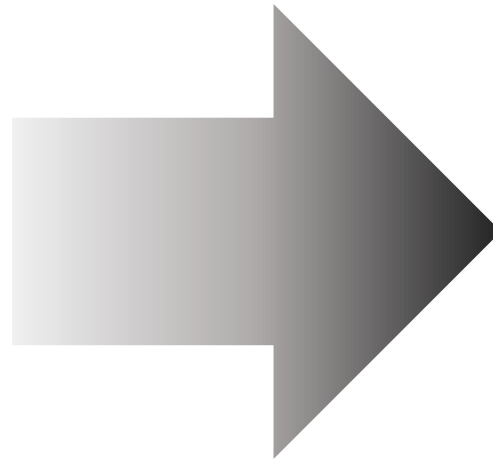
분석 활용 데이터를 이용하여 산사태 예측 모형 구축 시나리오

기상 데이터

- (AWS) AWS번호, 강수량
- (파생변수) 일강수량, 1일~3일 전 누적강수량, 시간당강수량, 최대강우강도, 최대순간풍속

비기상 데이터

- 산사태 발생이력
- (지형공간인자) 임상도, 토양도, 임도
- 행정구역 경계



산사태 발생 영향 변수 선택

정의	회귀계수가 높은 변수를 기준으로 순서대로 상관계수가 0.45이상인 변수를 제거하여 변수를 선택
분석방법	<ul style="list-style-type: none"> • 피어슨 상관분석 • 단계적 변수선택법
분석결과	산사태 발생 영향 변수 선택



산사태 발생 예측 모형 구축

정의	여러 모형을 이용하여 산사태 발생과 지형공간 및 기상인자 간의 관계를 나타내는 식 도출
분석방법	<ul style="list-style-type: none"> • 로지스틱 회귀모형 • 트리 기반 모형
분석결과	산사태 발생 유무 예측 모형

모형 구축

모형 적합 시 고려해야 할 점

	event	rn0	rn1	rn2	rn3	rn_tim	rn_hr1	wd
1	0	475.25	997.25	1056.75	1060.25	20.50	17.00	7.45
2	0	307.25	307.25	307.25	307.25	7.50	6.75	7.20
3	0	339.25	339.25	340.50	420.25	8.75	8.00	8.60
4	0	4.00	4.00	4.00	24.50	4.00	4.00	9.60
5	0	429.75	846.00	846.00	846.00	12.25	11.25	9.20
6	0	121.50	135.00	216.75	216.75	16.00	16.00	13.70
7	0	296.50	371.75	371.75	376.25	12.50	11.50	7.10
8	0	1085.25	1086.00	1096.25	1125.00	13.50	12.00	10.10
9	1	1468.25	2553.50	2554.25	2564.50	31.50	24.75	17.75
10	0	1058.85	1652.20	1652.20	1652.20	17.50	16.15	12.75
11	0	162.30	294.10	294.10	294.10	6.10	5.30	10.10
12	0	29.90	29.90	29.90	29.90	4.60	4.35	10.25

- 단계적 선택법을 이용하여 로지스틱 회귀분석을 시행한 결과, 산사태 발생에 유의한 변수로 기상변수들이 도출
- 기상변수들만을 사용하여 모형들을 적합했을 때, overfitting의 문제가 빈번히 발생

문제점 및 원인

- 기상 변수만을 사용하여 모형을 적합하였을 때, 모형이 조금만 복잡해져도 overfitting 발생
- 학습시간도 조금만 길어질수록 overfitting 발생

해결 방안

- 이를 해결하기 위해, 임상 및 토양 변수를 추가하여 overfitting 문제를 해결하고자 함
- 최대한 간단하고 기본적인 모형을 선택하고자 함

모형 구축

임상, 토양 정보 추가하면서 발생한 문제들



문제점 및 원인

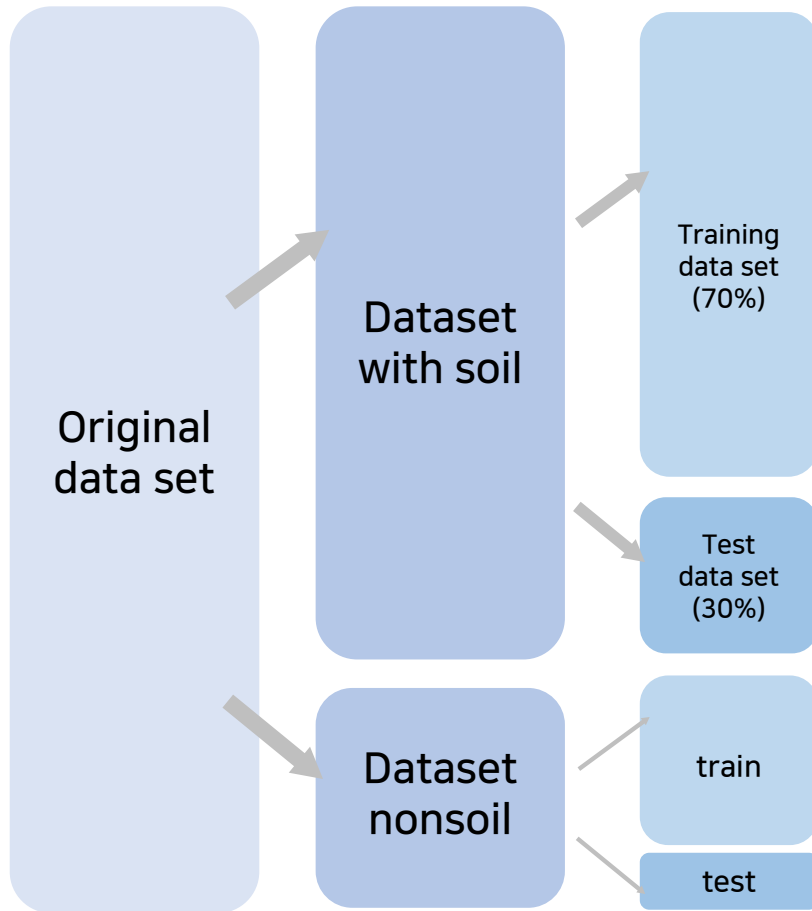
- 토양 변수를 추가하면서 데이터 일부에 공백이 있는 것을 확인
- 산림 공간 정보를 이용하여 적합시키는 경우에도 조금만 조합을 해도 overfitting 발생

해결 방안

- 이를 해결하기 위해, 산림 공간 정보의 유무를 기준으로 데이터셋을 구분
- 구분한 데이터에 대해 최대한 간단한 모형 적합을 시도하였으며, 각각 조합을 검토하여 최적의 조합을 찾고자 함
- 향후 이 모형을 통해 해석 및 사용하기 위해서 최근 유행하는 모형이 아닌, 데이터를 세분화하여 세분화된 데이터 별로 각각 모형을 적합시켜 계층적 모형을 통해 해결하고자 함

데이터 파악

train dataset



	sum_cnt	LOCTN_ALTT	LOCTN_GRDN	EIGHT_AGL	DNST_CD_C	DNST_CD_B	DNST_CD_A	AGCLS_CD_9
1	1	135.62837	19.17700	177.6269	0.7552654	0.05918281	0.008635215	0.000000000
2	1	114.31826	24.67776	181.7684	0.8393912	0.05270332	0.005452067	0.000000000
3	0	268.25300	25.44975	168.8182	0.9267016	0.01430023	0.024537001	0.000781433
4	0	116.76439	20.45145	163.8859	0.7458537	0.15743902	0.054634146	0.000121951
5	1	409.69636	25.22503	179.6610	0.8105399	0.09709294	0.037376486	0.000000000
6	0	423.11529	27.10160	184.6140	0.7185557	0.16021043	0.019846963	0.000000000
7	1	244.19007	23.82058	171.6145	0.9046654	0.03160454	0.018870109	0.001331327
8	0	173.31762	21.05222	151.3618	0.8243156	0.11384863	0.009017713	0.000000000
9	0	588.66376	27.18340	186.6394	0.7837174	0.09677419	0.037890425	0.000000000

	sum_cnt	LOCTN_ALTT	LOCTN_GRDN	EIGHT_AGL	DNST_CD_C	DNST_CD_B	DNST_CD_A	AGCLS_CD_9
1	1	102.42254	18.03963	169.4220	0.8162240	0.06712132	0.004666188	0.000000000
2	1	102.42254	18.03963	169.4220	0.8162240	0.06712132	0.004666188	0.000000000
3	1	108.95942	15.97601	159.7763	0.7312395	0.17285474	0.020863713	0.000000000
4	0	108.95942	15.97601	159.7763	0.7312395	0.17285474	0.020863713	0.000000000

	sum_cnt	rn_0	rn_1	rn_3	max
1	0	0.000000	1.900000	227.40000	0.0000000
2	1	3200.500000	3200.500000	3201.00000	70.5000000
3	0	227.966667	1528.433333	1781.13333	9.5000000
4	0	0.000000	0.000000	21.50000	0.0000000

	sum_cnt	rn_0	rn_1	rn_3	max
1	0	0.0	0.0	1.50000	0.0
2	0	0.0	1433.0	2302.00000	0.0

모형 성능

분류 성능 평가 지표

평가 지표

- **정분류율(ACC; Accuracy)** : 산사태가 일어난 상황에서 1)산사태가 발생한 경우, 2)산사태가 발생하지 않은 경우, 산사태가 일어나지 않은 상황에서 3)산사태가 발생한 경우, 4)산사태가 발생하지 않은 경우 등 총 4가지를 모두 조합해서 예측이 얼마나 정확한지를 계산
⇒ 즉, 산사태가 일어나는 상황과 산사태가 일어나지 않는 상황 모두를 고려하여 정확도를 계산
- **임계성공지수(CSI; Critical Success Index)** : 산사태가 발생하지 않는다는 상황에서 산사태가 발생하지 않는 경우의 값은 예보로써의 가치가 낮다는 입장이 반영된 지표
⇒ 즉, 산사태가 발생한다는 예측이 얼마나 맞았는지를 확률로 계산한 값

모형 구축

분류 분석

간편하고 해석에 용이
약한 예측력

로지스틱
회귀분석

의사결정나무

트리 기반 모형 & 앙상블 기법
변수 중요도 확인 및 강한 예측력

배깅


부스팅

랜덤포레스트

XGBoost

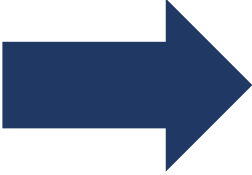
모형 검증

train data 분석 결과에 대한 비교

train_with_soil		validation
Logistic regression		ACC: 0.9576, CSI: 0.2314, AUC: 0.9474
Decision tree		ACC: 0.9602, CSI: 0.2308, AUC: 0.9482
Bagging		ACC: 0.9674, CSI: 0.2386, AUC: 0.9635
Boosting		ACC: 0.9870, CSI: 0.2425, AUC: 0.9817
Random Forest		ACC: 0.9824, CSI: 0.2419, AUC: 0.9778
XGBoost		ACC: 0.9831, CSI: 0.2425, AUC: 0.9791

모형 검증

train data 분석 결과에 대한 비교

train_non_soil		validation
Logistic regression		ACC: 0.8889, CSI: 0.2222, AUC: 0.9
Boosting		ACC: 0.8889, CSI: 0.2222, AUC: 0.9
Random Forest		ACC: 1.0, CSI: 0.2222, AUC: 1.0
XGBoost		ACC: 0.8889, CSI: 0.2222, AUC: 0.9

모형 비교

분류 분석 결과에 대한 비교

train_with_soil

- validation dataset에서의 최대 CSI 값은 0.2497로 4가지 모형 모두 높은 지표를 가짐
- 전반적으로 높은 ACC와 CSI 값을 가짐
- 앙상블 기법을 이용하여 적합한 모형이 예측력이 높았지만, 해석력 측면에서 강점을 보이는 로지스틱 회귀모형도 최종 모형의 후보로 결정

train_non_soil

- 랜덤 포레스트를 제외한 나머지 분류 기법 간의 큰 차이가 없는 것을 볼 수 있음
- 모형 적합 시 사용한 데이터의 개수가 22개로 비교적 부족

모형 검증

test data 분석 결과에 대한 비교

test_with_soil		test_non_soil		validation
Random Forest	+	Random Forest	=	ACC: 0.9152, CSI: 0.2340
Random Forest		Boosting		ACC: 0.9150, CSI: 0.2316
Logistic regression		Random Forest		ACC: 0.9077, CSI: 0.2297
Random Forest		Logistic regression		ACC: 0.9073, CSI: 0.2162
Logistic regression		Logistic regression		ACC: 0.9008, CSI: 0.2166
Boosting		Random Forest		ACC: 0.9152, CSI: 0.1931
XGBoost		Random Forest		ACC: 0.8704, CSI: 0.1470

최종 모형 선택

최종 모형을 이용하여 test data 예측

test_with_soil

- (model) Random Forest
- (parameter)
ntree = 1,000
mtry = sqrt(#obs_train_with_soil)
≈ 8.774964

test_non_soil

- (model) Random Forest
- (parameter)
ntree = 1,000
mtry = sqrt(#obs_train_non_soil)
≈ 2.828427

date	sd	sgg	umd	1day	2day	day1_yn.x	day2_yn.x
2020-07-10	경상남도	거제시	삼거동	2020-07-11	2020-07-12	0	0
2020-07-11	경상남도	거제시	삼거동	2020-07-12	2020-07-13	0	1
2020-07-12	경상남도	거제시	삼거동	2020-07-13	2020-07-14	1	0
2020-07-13	경상남도	거제시	삼거동	2020-07-14	2020-07-15	0	0
2020-08-06	경상남도	거제시	삼거동	2020-08-07	2020-08-08	0	0

date	sd	sgg	umd	1day	2day	day1_yn.y	day2_yn.y
2020-09-04	경상북도	울릉군	북면	2020-09-05	2020-09-06	0	0
2020-09-05	경상북도	울릉군	북면	2020-09-06	2020-09-07	0	0
2020-09-06	경상북도	울릉군	북면	2020-09-07	2020-09-08	0	0
2020-08-06	경상북도	울릉군	울릉읍	2020-08-07	2020-08-08	0	0
2020-08-07	경상북도	울릉군	울릉읍	2020-08-08	2020-08-09	0	0
2020-08-08	경상북도	울릉군	울릉읍	2020-08-09	2020-08-10	0	0

최종 모형

분류 분석 결과에 대한 최종 모형

결론

- 분석 결과, 기상인자가 산사태의 발생에 주요한 영향을 미치는 것으로 확인
- ACC는 모형 간 큰 차이 나지 않지만, CSI를 기준으로 봤을 때 가장 큰 값을 가지는 랜덤 포레스트 모형을 최종 모형으로 선택
- 최종 모형의 예측정확도는 0.9152, 임계성공지수(CSI)는 0.234로 높은 것으로 나타남

한계

- 일부 지역의 임상도 및 토양도 데이터 마스킹으로 인한 자료 구축의 한계로 산사태 발생 지역의 산림 공간 정보의 반영 부족
- 산사태와 같은 자연재해의 발생을 정확히 예측하기 위해서는 시·공간적인 분석이 필요하고, 지형 공간인자와 기상인자 등 성격이 다른 인자들 사이의 역학 관계와 기작 등을 파악하는 것이 필요

결론 및 제언

결과 활용도

제언

- 데이터의 질 측면에서 시간과 공간의 시점이 동일하지 않아 정확한 분석이 불가능한 점이 아쉬움
- 산사태 예측에 영향을 주는 임상 및 토양 변수를 확인할 수 있었고, 토양의 경사보다 토양의 종류와 임상 정보가 영향을 주는 것을 보아 산사태 방지를 위해 새로운 임상을 복구할 때 어떤 나무를 심을지, 이때 어떤 토양을 사용할지 등 제안할 수 있음

시사점

- 산사태 발생에 기상요소가 많은 영향을 미치는 것은 알고 있지만, 연구 진행 과정에서 기상변수만 사용할 경우 대부분 모형에서 overfitting이 발생하였음
- 기상변수와 임상변수를 모두 사용할 경우에도 조금만 모형이 복잡해지면 바로 overfitting이 발생하는 것을 확인할 수 있음
- 즉, 산사태 발생에는 여러 정보가 복합적으로 영향을 끼치므로 정확한 연구를 위해서는 여러 분야를 아우르는 복합적인 연구가 필요해 보임



기상청

감사합니다.

