발 간 등 록 번 호
11-1360000-001381-01

기상청 기술노트

2016년 기상청 內 빅데이터 분석 서비스

- 관측(증발량, 운량), 항공 -

기상청

정부3.0

2016년 기상청 內
빅데이터 분석 서비스

- 관측(증발량, 운량), 항공 -

기상융합서비스과장 **오 미 림**
방송통신사무관 **김 진 석**
기상주사 **경 규 정**
기상주사보 **조 은 주**
연구원 **박 수 현**

〈차　　례〉

〈제목 차례〉 ····· i
〈표 차례〉 ····· ii
〈그림 차례〉 ····· iii
1. 서　론 ····· 1
1.1. 청 내 빅데이터 분석 서비스 ····· 1
1.1.1. 개요 ····· 1
1.1.2. 청 내 빅데이터 분석 서비스 지원 절차 ····· 2
1.1.3. 빅데이터 분석기법 ····· 2
2. 2016년 청 내 빅데이터 분석 서비스 ····· 5
2.1. 증발량 산출식 개발 ····· 5
2.1.1. 개요 ····· 5
2.1.2. 데이터 수집 및 탐색 ····· 6
2.1.3. 데이터 전처리 ····· 8
2.1.4. 분석 실행 및 결과 ····· 12
2.1.5. 검증 결과 및 결론 ····· 15
2.2. 운량 산출식 개발 ····· 21
2.2.1. 개요 ····· 21
2.2.2. 천리안 위성 및 운고운량계 분석 ····· 22
2.2.3. 히마와리 위성 및 관측 정보 분석 ····· 28
2.2.4. 검증 결과 및 결론 ····· 32
2.3. 항공기 출발 지연 분석 ····· 37
2.3.1. 개요 ····· 37
2.3.2. 데이터 수집 및 탐색 ····· 38
2.3.3. 데이터 전처리 ····· 40

2.3.4. 분석 실행 및 결과 ····· 41
2.3.5. 항공기 출발 지연 위험도 임계치 산출 ····· 42
2.3.6. 결론 ····· 47
부록 ····· 51
부록 1. 증발량 산출식 개발 ····· 51
부록 2. 운량 산출식 개발 ····· 52
부록 3. 항공기 출발 지연 분석 ····· 54
참고문헌 ····· 55
찾아보기 ····· 56

〈표 차례〉

〔표 1.1〕 2016년 기상기후 빅데이터 분석 서비스 지원 절차 ····· 2
〔표 2.1〕 증발량 산출식 개발 수집 데이터 ····· 7
〔표 2.2〕 소형증발량과 대형증발량의 관계 ····· 9
〔표 2.3〕 증발량데이터의 로그변환(정규화) 전·후 히스토그램 비교 ····· 10
〔표 2.4〕 파생변수 생성에 활용된 PM산출식 변수의 공식 ····· 11
〔표 2.5〕 문헌 연구 결과 생성된 파생변수 ····· 11
〔표 2.6〕 기상요소의 3시간 단위 파생변수 ····· 12
〔표 2.7〕 증발량 산출 모형별 최종 선택 변수 ····· 14
〔표 2.8〕 증발량 산출 모형의 성능 ····· 15
〔표 2.9〕 모형별 평균예측오차 비교 ····· 16
〔표 2.10〕 증발량 산출 모형의 지점별 예측오차(%) ····· 17
〔표 2.11〕 PM산출식과 증발량 산출 모형의 오차 평균 차이 ····· 17
〔표 3.1〕 운량 산출 분석별 결과 요약 ····· 22
〔표 3.2〕 천리안 위성 및 운고운량계 분석 수집 데이터 ····· 23
〔표 3.3〕 데이터별 결측치 탐색 결과 ····· 24
〔표 3.4〕 최종 분석 활용 변수 ····· 25
〔표 3.5〕 수집 데이터와 분석 대상 ····· 28
〔표 3.6〕 수집 데이터의 변수별 속성 ····· 29
〔표 3.7〕 데이터별 결측치 및 이상치 탐색 결과 ····· 30
〔표 3.8〕 최종 랜덤포레스트 모형 ····· 32
〔표 4.1〕 분석 대상 및 수집 데이터 ····· 38
〔표 4.2〕 수집 데이터 내 변수 ····· 39
〔표 4.3〕 최종 변수 선택 ····· 42
〔표 4.4〕 항공기 출발 지연 예측 모형 결과 ····· 43
〔표 4.5〕 항공기 출발 지연에 영향을 주는 기상변수들의 위험도 임계치 ····· 45
〔표 4.6〕 공항별 출발 지연 확률 및 위험 등급 ····· 46

〈그림 차례〉

〔그림 2.1〕 증발량 산출식 개발 분석 과제 요약 ····· 6
〔그림 2.2〕 증발량 데이터 요약 ····· 8
〔그림 2.3〕 증발량 산출식 개발 위한 분석 데이터 셋 ····· 9
〔그림 2.4〕 증발량과 기상변수의 시간 해상도 매칭 ····· 10
〔그림 2.5〕 증발량 산출식 개발 분석 절차 ····· 12
〔그림 2.6〕 시간별 최대 일사량과 소형증발량의 관계 ····· 15
〔그림 3.1〕 천리안 위성 및 운고운량계 분석 절차 ····· 24
〔그림 3.2〕 분석 모형 정의 ····· 26
〔그림 3.3〕 분석 모형 및 추정 회귀계수 ····· 26
〔그림 3.4〕 운량 산출 모형의 예측 적합도 ····· 27
〔그림 3.5〕 관측소별 운량 산출 모형 검증 ····· 27
〔그림 3.6〕 전운량 분포 ····· 29
〔그림 3.7〕 히마와리 위성 및 관측 정보 분석 절차 ····· 31
〔그림 3.8〕 운량 산출 모형의 예측 적합도 ····· 32
〔그림 3.9〕 관측소별 운량 산출 모형 검증 ····· 33
〔그림 4.1〕 항공기 출발 지연 분석 과제 요약 ····· 37
〔그림 4.2〕 비정상 운항 정의를 위한 운항 데이터 탐색 ····· 39
〔그림 4.3〕 결측치 및 이상치 처리 기준 ····· 40
〔그림 4.4〕 기상파생변수 생성 ····· 41
〔그림 4.5〕 항공기 출발 지연 분석 데이터 셋 ····· 41
〔그림 4.6〕 항공기 출발 지연 분석 절차 ····· 42
〔그림 4.7〕 일기코드별 출발 지연 확률 및 위험 등급 ····· 31

1. 서론

1.1. 기상청 내 빅데이터 분석 서비스

1.1.1. 개요

무선 인터넷과 스마트 폰을 필두로 한 모바일 기술이 일상화되고 동시에 데이터 저장·처리를 위한 IT 비용이 급격하게 하락하면서 최근 우리 주위에 데이터가 넘쳐나고 있다. 바야흐로 데이터 폭발의 시대, 빅데이터의 시대가 도래한 것이다.

> Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, and process automation. - Gartner, Inc.-

미국 IT 자문 회사인 가트너(Gartner, Inc.)에 의해 널리 알려진 '빅데이터'의 정의는 높은 통찰력, 의사결정 프로세스 자동화를 위해 비용효과가 높은 혁신적인 정보처리 과정을 요하며, 대용량의 데이터 규모, 빠른 속도, 다양성이 높은 정보 자산이다. 이를 간단히 3V(Volume, Velocity, Variety)라고 하며 최근 가치(Value), 정확성(Veracity) 등 지속적으로 의미가 추가되고 있다. 또한 기술과 IT 생태계가 발전함에 따라 대용량 데이터를 활용 및 분석하여 가치 있는 정보를 추출하고, 생성된 정보를 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 기술의 집합으로서 포괄적인 의미를 담고 있다.

기상청은 입체관측망을 통해 수집된 관측자료와 세계기상자료 등 대량의 자료를 분석하여 내일의 날씨 변화를 매일매일 예측·생산하고 이를 국민들에게 제공하여 왔다. 이런 측면에서 기상청은 빅데이터가 화두가 되기 훨씬 이전부터 이미 빅데이터를 충분히 활용하고 있다고 하겠다.

기상정보는 과학을 기반으로 산출되고, 개인정보 보호에 대한 이슈가 없어 데이터 활용에 제약이 낮은 대표적인 공공 데이터다. 기상청이 보유한 대량의 데이터가 타분야 융합을 통해 예보 생산을 넘어, 청 내의 업무 개선, 새로운 기상기후 서비스 개발 등에 활용된다면 기상청의 발전에 도움이 될 것이다.

기상융합서비스과는 청내 현안 해결 및 데이터 중심의 기상행정 구현을 위하여 과거 수 십 년 전부터 수집된 관측·예보자료, 수치모델자료와 같은 대용량 기상자료와 해발고도, 산림면적, 항공 운항 정보 등 비기상 자료를 융합하여 청 내 빅데이터 분석 과제를 지원하고 있다.

본 기술노트에서는 2016년 관측과 항공기상 분야의 청 내 빅데이터 분석 서비스에 대해 구체적으로 기술하고 있다.

1.1.2. 청 내 빅데이터 분석 서비스 지원 절차

빅데이터에 대한 관심과 빅데이터 분석을 통한 데이터 기반의 정책 수립 강화의 필요성이 증가하고 있다. 이에, 청 내 현안 해결 및 데이터 중심의 기상행정을 구현하고자 2016년 2월 청 내 빅데이터 분석 과제 수요조사를 시행하였다. 기상 업무 전체 분야(일반행정, 관측, 예보, 기후, 서비스 등) 또는 기상과 타 분야의 데이터 융합을 통한 새로운 기상기후 서비스 발굴 및 개발 분야에 대하여 수요조사를 하였다. 제출된 과제 중 업무 개선과 효율화 또는 신규 사업 발굴을 위한 사전 검증을 목표로 하는 과제를 선정하여 분석을 수행하였으며, 분석 시 기계학습 기반의 분석기법을 활용하였다. 그리고 그 결과를 수요부서에 환류하였다.

[표 1.1] 2016년 기상기후 빅데이터 분석 서비스 지원 절차

분석 수요조사 (수요부서)	⇒	과제선정 (자체심의회구성)	⇒	과제수행 (기상융합서비스과, 내외부전문가)	⇒	분석결과활용 (수요부서)

1.1.3. 빅데이터 분석기법

과거 데이터 분석의 목적이 현황, 인과 관계 파악이라면 최근 대두되고 있는 빅데이터 분석은 미래 현상 예측, 의사결정 지원을 위한 새로운 가치를 발굴하는 것이 목적이다. 사실을 내포하고 있는 동시에 미래를 말하고 있는 데이터의 분석을 통해 어떤 현상에 대한 법칙 등을 발견하고 통찰력을 찾는 과정이 빅데이터 분석이라 할 수 있다.

빅데이터 분석기법에는 전통적인 통계분석(상관분석, 회귀분석 등)부터 기계학습 기법(랜덤포레스트, 딥러닝 등)까지 분석주제, 데이터 형태, 분석 대상에 따라 다양한 유형의 분석기법이 존재한다. 즉, 이전에는 없었던 새로운 분석기법이 아니라 기존의 분석기법부터 성능을 고도화 시킨 고급분석까지 모두를 아우르는 것이라 할 수 있다. 선형분석은 모형이 단순하여 해석에 용이하고, 요인에 따른 민감도를 파악 가능하지만 예측모형이 극단값 등에 민감하게 반응하는 단점이 있다. 비선형분석은 모형이 복잡하여 해석이 어렵고, 요인의 관계를 파악하기 어렵지만 모형이 안정적으로 예측력이 좋은 장점이 있다. 대부분의 수요자는 분석대상에 영향을 미치는 요인과 그 요인별 민감도 및 예측력을 알고자 한다. 하지만, 모두를 충족시키는 결과를 얻는 것은 하늘의 별따기이다.

2016년 청 내 빅데이터 분석 서비스 지원에 활용된 분석기법으로는 기계학습 기법인 랜덤포레스트(Random Forest)와 일반화선형모형(General Linear Model) 등을 이용하였다. 분석기법에 대한 설명은 본 기술노트 내 본문에 정리되어 있다.

청 내 빅데이터 분석 서비스

(관측) 증발량 산출식 개발

2. 2016년 청 내 빅데이터 분석서비스

2.1. 증발량 산출식 개발

2.1.1. 개요

1964년부터 기상법 제7조 및 관측업무규정 제15조(기후관측), 제21조(농업기상관측)에 따라 기후 및 농업기상관측관서[1]에서 일 1회 9시에 증발량 관측을 수행 중이다. 증발량은 직접 관측이 어렵고, 겨울에는 물이 어는 현상 등 여러 오차 요인들이 많아 측정의 정확성이 매우 낮은 편으로 활용도도 낮고, 관측하지 않는 나라들이 늘어나고 있는 실정이다.

◆ 전일근무관서 22개소 중 기후·농업관서 16개소 증발량 관측
- 기후관서(11개소) : 서울, 부산, 포항, 여수, 목포, 대전, 인천, 강릉, 춘천, 울릉도, 제주
- 농업관서(6개소) : 수원, 안동, 전주, 청주, 춘천, 서산(철원, 순천, 진주, 서귀포 미수행)

◆ 증발량과 증발산량 정의
- 증발량 : 개방된 물 표면이나 지면으로부터 증발된 물의 양
- 증발산량 : 지면이 자연 상태의 습기를 가지고 있을 때 토양과 식물로부터 증발된 수증기의 양

해외의 경우 증발량 목측을 중지한 나라가 있는 반면, 유지하고 있는 나라도 있다. 먼저, 증발량 목측을 중지한 영국, 독일, 중국, 일본 등의 경우, 증발량을 산출하는 방법으로 전환하였다. 영국은 MORECS(Met Office Rainfall and Evaporation Calculation System)에서 증발량 산출하고 있으며, 독일은 Penmann-Monteith 방법과 토양수분 모델(AMBAV)에서 증발량 산출하고 있다. 프랑스는 증발량 관측하지 않으며, 잠재 증발량을 산출하고 있고, 중국은 2002년부터 자동관측(대형증발계+초음파식 거리측정기, E601B 모델)을 활용하고 있다[2]. 그 밖의 스위스 등도 목측은 하지 않고 있다.

증발량 목측을 유지하고 있는 나라는 미국, 호주 등이 있는데, 미국의 경우 협력(위탁) 관측을 하고 있으며 호주의 대부분 또한 협력(위탁) 관측 중이나, Penmann-Monteith 방법 또는 자동관측(일부 지점)으로 변경을 계획 중인 것으로 알려져 있다.

이에 기상청에서는 일반적인 자동화된 기상관측요소로부터 산출될 수 있는 증발량 산출 관계식을 빅데이터 기법을 통하여 개발하고자 하였다. 자동화된 기상관측 요소로부터 증발량을 산출할 수 있다면 증발량을 직접 관측하지 않는 관서 및 AWS 등에도 확대 적용이 가능할 것이며, 관측 업무 개선, 증발량 검증 및 진단식으로 활용할 수 있을 것으로 기대되어 본 분석을 시행하게 되었다.

---

1) 기상청 농업기상관측관서(11개소) : 수원, 청주, 철원, 춘천, 서산, 안동, 전주, 진주, 순천, 서귀포, 보성군
2) 중국은 1951~2002년 소형증발계(직경 20cm), 1998년 이후 대형증발계(Russian GGI-3000 pan과 유사) 사용, 2002년부터 자동관측 실시(Xiong et al., 2012: Reconstruction of a Daily Large-Pan Evaporation Dataset over China. Journal of Applied Meteorology and Climatology.)

따라서, 2016년 청 내 빅데이터 분석 서비스 중 증발량 과제는 ASOS(지상기상관측) 관측요소에 빅데이터 기법을 적용하여 증발량을 산출할 수 있는 관계식을 도출하는 것을 목적으로 하였으며, 일반선형모형(GLM, General Linear Model)과 랜덤포레스트(Random Forest)를 통한 산출 결과와 Penmann-Monteith 산출식을 통한 증발량 산출 결과를 비교하였다. 증발량 산출식 개발 분석 과제에 대한 요약은 다음 [그림 2.1]과 같다.

➢ **Penmann-Monteith 일별 증발산량, 증발량 계산식 (PM산출식)**

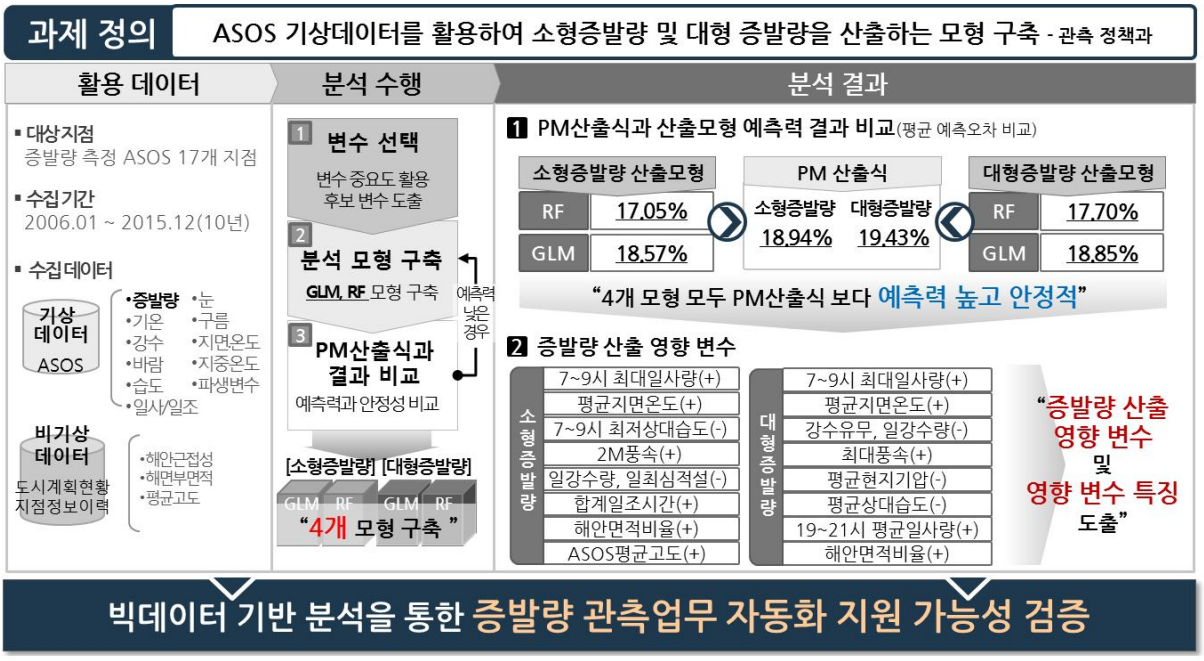
- 증발산량 산출식 : $ET_o = \dfrac{0.408\Delta(R_n - G) + \gamma \dfrac{C_n}{T+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + C_d u_2)}$

  ET_o는 증발산량(evaporation, mm/hr)
  R_n는 순복사량(net radiation, $MJ/m^2/day$)
  G는 토양열속밀도(soil heat flux density, $MJ/m^2/day$) ≈ 0 (for day)
  γ는 건습계 상수(psychrometric constant, kPa/C)
  Δ는 기온에 따른 포화수증기압의 기울기(kPa/C)
  u_2는 2m 높이에 해당하는 일평균 풍속(m/s)
  e_s는 일평균 포화수증기압(kPa)
  e_a는 일평균 수증기압(kPa)
  T는 일평균 기온(C), u_2는 2m 높이에 해당하는 일평균 풍속(m/s)
  C_n과 C_d는 작물의 종류에 따른 계수(잔디인 경우, $C_n = 900$, $C_d = 0.34$)

- 증발량 계산 : $EV = ET_o / K_p$

  EV는 증발량(mm/day), ET_o는 증발산량(mm/day)
  K_p는 팬 계수로 잔디의 경우 아래 식과 같으며, FET는 지점별 특성값으로 100으로 한다.

$$K_p = 0.108 - 0.0286u_2 + 0.0422\ln(FET) + 0.1434\ln(RH_{mean}) - 0.000631\,[\ln(FET)]^2 \ln(RH_{mean})$$



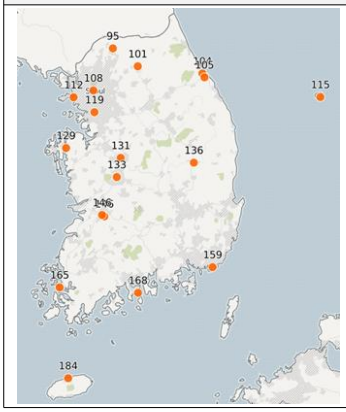
[그림 2.1] 증발량 산출식 개발 분석 과제 요약

2.1.2. 데이터 수집 및 탐색

❍ 데이터 수집 및 분석 대상 선정

증발량 산출식 개발을 위해 수집된 데이터 목록은 [표 2.1]과 같다. 분석 대상 기간은 2006년 1월 1일부터 2015년 12월 31일까지 10년간이며, 현재 증발량을 관측하고 있거나 증발량 관측 데이터가 확보된 17개 관측소를 대상으로 하였다. 이 때, 강릉 관측소의 경우 2008년 8월 관측이 종료되었으나 북강릉 관측소에서 2008년 8월부터 관측이 시작되었으므로 강릉과 북강릉을 같은 관측소로 취급하여 분석하였다.

[표 2.1] 증발량 산출식 개발 수집 데이터

기상기후 데이터	비(非)기상 데이터
ASOS	지형데이터
수집기간 : (증발량) 2006.01～2015.12(10년, 1일단위) (기상정보) 2006.01～2015.12(10년, 1시간단위) 지점 관측소 : 증발량 관측 17지점 데이터건수 : (증발량) 58,023 (기상정보) 62,084 변수종류 : 소형증발량, 대형증발량, 기온, 강수량, 풍속, 습도, 일조량, 기압, 안개, 증발량	수집기간 : 2006.01.01. ~ 2015.12.31. 지점 관측소 : 증발량 관측 17지점 데이터건수 : 17건 변수종류 : ASOS 해발고도, 해면부면적 출처 : 통계청, 기상청
증발량 분석 대상 관측소	일출/일몰시간 데이터
 95(철원), 101(춘천), 104(북강릉), 105(강릉), 108(서울), 112(인천), 115(울릉도), 119(수원), 129(서산), 131(청주), 133(대전), 136(안동), 146(전주), 159(부산), 165(목포), 168(여수), 184(제주)	수집기간 : 2006.01.01. ~ 2015.12.31. 지점 관측소 : 증발량 관측 17지점 데이터건수 : 58,051건 변수종류 : 일출 시간, 일몰 시간 (가조시간 계산을 위해 수집) 출처 : KASI(한국천문연구원)

❍ 데이터 탐색 및 결측치[3]·이상치[4] 처리

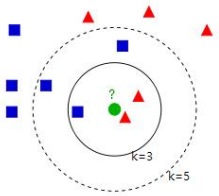
데이터 탐색은 데이터 현황 및 특성을 파악하고, 데이터 전처리 과정을 통해 분석 목적에 적합한 데이터 셋을 정의하기 위한 선행과정이다.

증발량 외의 기상 데이터의 이상치는 기후자료 품질검사 알고리즘 기준(부록 1-1 참고)으로 판별하였으며 결측치는 kNN Imputation 방법을 활용하여 대체하였다.

3) 결측치(Missing value) : 관측 장비 장애, 미관측 등의 이유로 값이 없는 데이터
4) 이상치(Outlier) : 값은 존재하지만 비정상적인 값의 데이터

➢ **kNN(k Nearest Neighbor)-Imputation**

◆ 데이터 결측치 존재 시, 예측하고자 하는 데이터로부터 가장 가까운 k개의 이웃을 찾은 뒤 이들 이웃으로부터 예측하고자하는 데이터의 분류를 정하여 값을 대체하는 방법

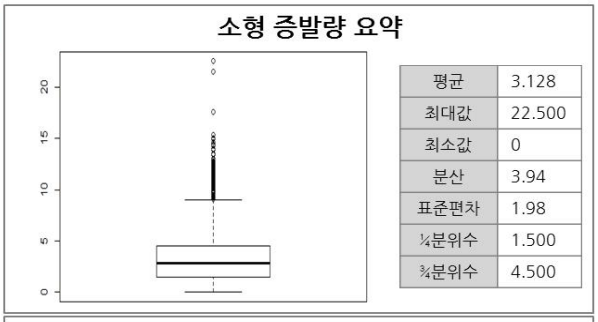


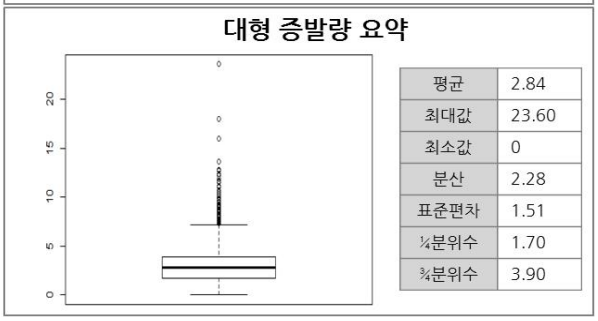
증발량 데이터 요약		
수집 대상	증발량 측정 ASOS 17개 지점[1]	
수집 기간	2006년 1월 1일~2015년 12월 31일	
총 데이터 수	58,023 건	
대형증발량	관측치(미관측치[2])	34,005 건(24,018건)
	결측치	91 건(0.267%)
	이상치[3]	3건(0.005%)
소형증발량	관측치	58,023 건
	결측치	2건(0.003%)
	이상치[3]	5건(0.009%)

1) 총 17개 지점 (95, 101, 104, 105, 108, 112, 115, 119, 129, 131, 133, 136, 146, 159, 165, 168, 184)

2) 11~3월 : 대형 증발량 미관측 기간

3) 이상치는 음수 또는 15를 초과하는 값을 기준으로 함 (*기후자료 품질검사 알고리즘 기준값)





[그림 2.2] 증발량 데이터 요약

증발량 데이터의 이상치 판단은 [그림 2.2] 우측의 Box plot[5]과 전문가의 의견을 토대로 음수 또는 15mm를 초과하는 값을 기준으로 하였다. 소형증발량 데이터의 경우, 결측치 및 이상치 데이터가 58,023건 중 7건으로 거의 없다고 할 수 있을 정도로 미미하였다. 대형증발량 데이터의 경우, 매년 11월부터 다음해 3월까지 미관측 기간이 존재하여 총 34,005건으로 소형증발량 데이터보다 동일 기간 내 데이터의 수가 작고, 총 94건이 결측치 및 이상치로 분류되었다. 이와 같이 분류된 결측치·이상치 데이터는 분석에서 제외하였다.

❍ 소형증발량과 대형증발량의 관계

먼저, Pearson 상관분석을 통하여 소형증발량과 대형증발량의 상관관계를 살펴보았다. [표 2.2]를 통해 알 수 있듯 평균적으로 0.954의 상관계수를 가지므로 소형증발량과 대형증발량은 매우 높은 양적 상관관계가 존재한다.

또한, 대형 증발량은 평균적으로 소형증발량의 0.7배임을 알 수 있다. 지점별로는 최고 0.74배(강릉, 서산, 대전), 최소 0.66배(북강릉, 울릉도, 제주) 차이를 보였다.

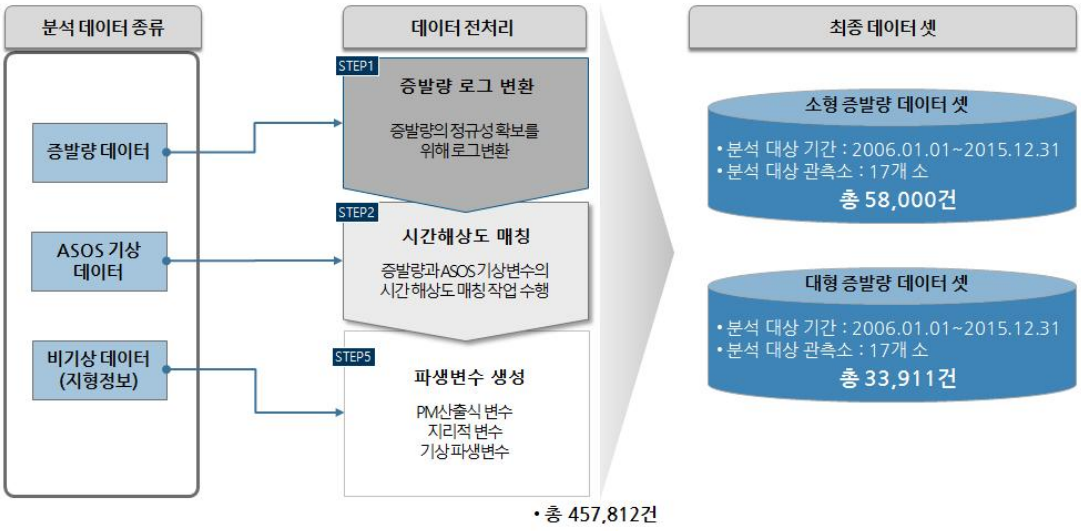
5) Box plot : 상자그림이라고도 하며, 최솟값, 1사분위, 중앙값, 3사분위, 최댓값을 통해 자료의 분포를 확인한다.

[표 2.2] 소형증발량과 대형증발량의 관계

지점번호	지점명	상관계수	대형증발량 ÷ 소형증발량
95	철원	0.951	0.67
101	춘천	0.948	0.70
104	북강릉	0.938	0.66
105	강릉	0.960	0.74
108	서울	0.935	0.69
112	인천	0.929	0.68
115	울릉도	0.909	0.66
119	수원	0.962	0.70
129	서산	0.973	0.74
131	청주	0.960	0.70
133	대전	0.975	0.74
136	안동	0.962	0.68
146	전주	0.964	0.71
159	부산	0.997	0.70
165	목포	0.967	0.73
168	여수	0.951	0.72
184	제주	0.944	0.66
평 균		0.954	0.70

2.1.3. 데이터 전처리

전처리 단계에서는 증발량 데이터의 정규성 확보를 위해 로그 변환을 하였으며, 증발량과 ASOS 기상 변수의 시간 해상도 매칭 작업을 하였다. 또한, 파생변수를 생성하여 분석에 활용하였다. 이를 요약한 내용은 [그림 2.3]과 같다.



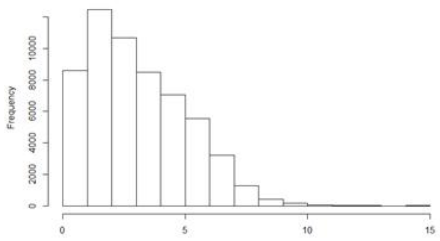
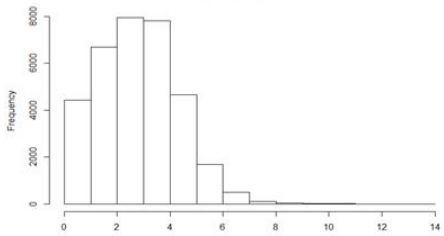
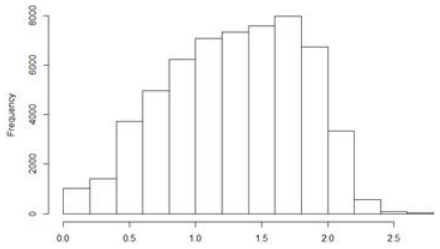
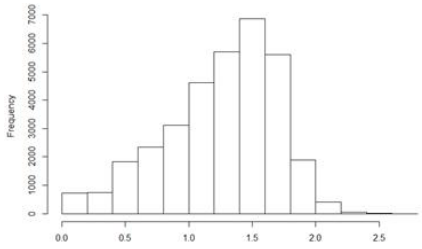
[그림 2.3] 증발량 산출식 개발 위한 분석 데이터 셋

❍ 데이터 변환

증발량 데이터의 히스토그램을 통해 분포를 살펴보면 0과 6사이에 밀집하고 있음을 알 수 있다. 데이터 중 '0'이 다수 존재하므로 모든 데이터에 1을 더하고, 밑이 e인 자연로그를 취하여 로그변환을 하였다. 제곱근 또는 Box-Cox 변환이 아닌 로그변환을 활용한 이유는 밀집된 데이터의 분포를 넓게 퍼트리기에 더 적합하기 때문이다. [표 2.3]의 히스토그램을 살펴보면 비교적 정규화 되었다고 할 수 있다.

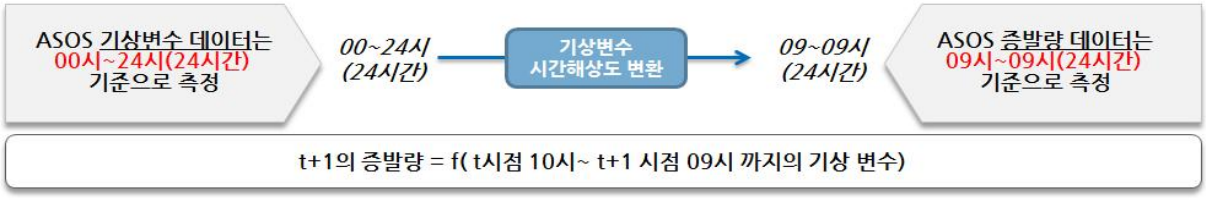
$$Y = \ln(X+1),\ where\ \begin{cases} Y\text{는 변환된 데이터} \\ X\text{는 원데이터} \end{cases}$$

[표 2.3] 증발량데이터의 로그변환(정규화) 전·후 히스토그램 비교

	소형증발량	대형증발량
원데이터		
로그변환된 데이터		

❍ 시간 해상도 매칭

ASOS 기상변수와 증발량 데이터의 시간 해상도 매칭을 위해 각 지점별 시간 해상도를 조정하였다. 증발량은 매일 오전 9시에 측정되므로 전날의 오전 9시 1분부터 다음날 오전 8시 59분까지의 영향을 받는다. 또한 ASOS 기상 데이터는 시간단위이므로 t+1시점의 증발량은 t시점의 오전 10시부터 t+1시점의 오전 9시까지의 기상변수와 결합된다.



[그림 2.4] 증발량과 기상변수의 시간 해상도 매칭

○ 파생변수 생성

분석 정확성 향상과 유의미한 정보를 찾기 위해 원시데이터로부터 여러 변수의 조합, 조정을 통해 의미 있는 파생변수를 생성한다.

- Penman-Monteith 증발량 계산식 활용

PM 증발량 계산식을 구성하는 다양한 요인들 중에서 ASOS 기상변수들로 생성 가능한 변수 4가지를 파생변수로 생성하여 활용하였다. '2m 높이에 해당하는 일평균 풍속', '일평균 포화수증기압', '기온에 따른 포화 수증기압의 기울기', '일평균 수증기압'을 생성하였으며 각 변수별 공식은 다음 [표 2.4]와 같다.

[표 2.4] 파생변수 생성에 활용된 PM산출식 변수의 공식

2m 높이에 해당하는 일평균 풍속(m/s)	일평균 포화수증기압(kPa)
$u_2 = u \times \dfrac{4.87}{\ln(67.8Ht - 5.42)}$ u는 풍속계 높이에서의 풍속(m/s) Ht는 풍속계의 지상 높이(m)	$e_s = \dfrac{e^o(T_{\max}) + e^o(T_{\min})}{2}$ 포화수증기압 계산식 : $e^o(T) = 0.6108\exp\left(\dfrac{17.27\,T}{T+237.3}\right)$ T는 기온(℃)
기온에 따른 포화 수증기압의 기울기(kPa/℃)	**일평균 수증기압(kPa)**
$\Delta = \dfrac{4098\left[0.6108\exp\left(\dfrac{17.27T}{T+237.3}\right)\right]}{(T+237.3)^2}$ T = 일평균 기온(℃) = $\dfrac{T_{\max} + T_{\min}}{2}$	$e_a = \dfrac{e^o(T_{\min})\dfrac{RH_{\max}}{100} + e^o(T_{\max})\dfrac{RH_{\min}}{100}}{2}$ RH는 상대습도(Relative Humidity)

- 문헌적 연구 결과를 통한 변수 활용

증발량 관련 문헌을 조사하고 해당 연구 결과를 토대로 증발량에 영향을 미치는 지리적 변수 3가지, 기상학적 변수 2가지를 생성하여 파생변수로 활용하였다.

[표 2.5] 문헌 연구 결과 생성된 파생변수

구분	변수 종류	변수 내용	특징
지리적 변수	해안근접성	해안지역(1), 내륙지역(0)	증발량의 지형적 영향력 반영
	해안부면적비율	해당지역의 해안부면적비율	
	평균고도자료	ASOS 평균고도	
기상학적 변수	입사태양복사량	대기권내에 도달하는 입사태양복사량	Precipitation, Temperature, Sunshine duration 계산식 활용 (가조시간, 일조시간, 강수량, 최저기온변수로 생성)
	강수유무	강수유무	0, 1

- 기상요소의 시간 단위 파생변수 활용

24시간을 3시간 단위로 나누어 각 시간별 기온, 풍속, 상대습도, 일사량의 최대, 최소, 평균값을 구하여 파생변수로 생성하였다. 총 80개의 파생변수를 생성하여 최종 데이터에 포함하였다.

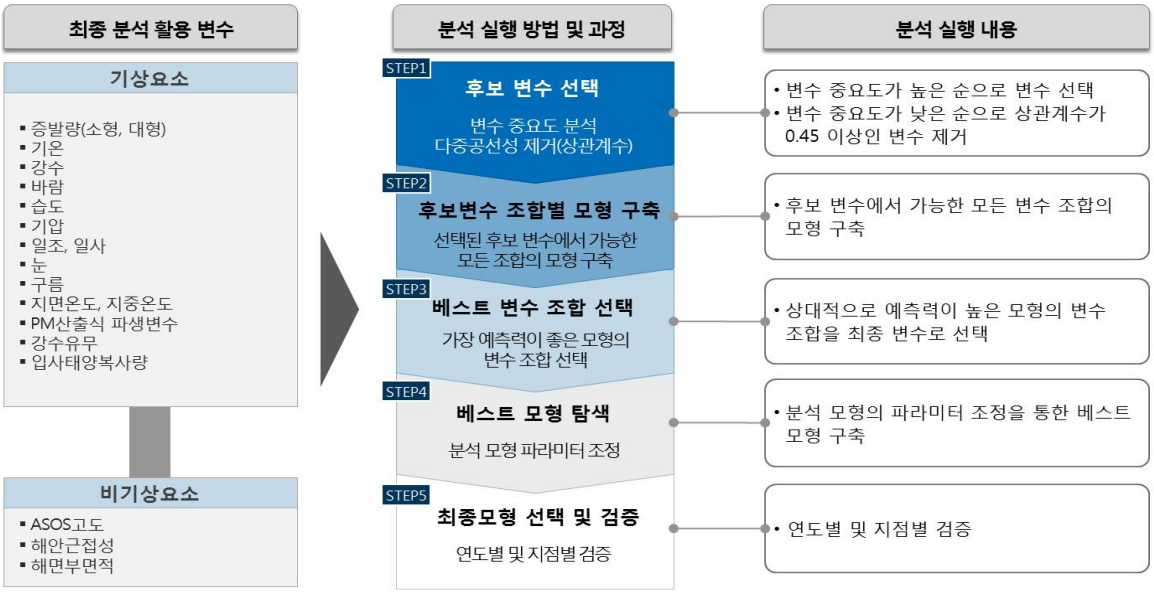
[표 2.6] 기상요소의 3시간 단위 파생변수

<table>
<tr><th>시간</th><th>10～12</th><th>13～15</th><th>16～18</th><th>19～21</th><th>22～24</th><th>01～03</th><th>04～06</th><th>07～09</th><th></th></tr>
<tr><td rowspan="2">기온</td><td colspan="8">최고 기온</td><td rowspan="10">총 80개
파생변수
생성</td></tr>
<tr><td colspan="8">평균 기온</td></tr>
<tr><td rowspan="3">풍속</td><td colspan="8">최대 풍속</td></tr>
<tr><td colspan="8">평균 풍속</td></tr>
<tr><td colspan="8">최저 풍속</td></tr>
<tr><td rowspan="3">상대
습도</td><td colspan="8">최고 상대습도</td></tr>
<tr><td colspan="8">평균 상대습도</td></tr>
<tr><td colspan="8">최저 상대습도</td></tr>
<tr><td rowspan="2">일사량</td><td colspan="8">최대 일사량</td></tr>
<tr><td colspan="8">평균 일사량</td></tr>
</table>

2.1.4. 분석 실행 및 결과

❍ 분석 실행 절차

2.1.3 데이터 전처리를 통해 정제된 112개 변수의 데이터 셋에서 변수 중요도와 상관계수를 고려하여 후보 변수를 선택하였다. 모형의 예측력을 고려하여 최종 모형을 구축하고 검증하였다. 자세한 분석 절차는 [그림 2.5]와 같다.



[그림 2.5] 증발량 산출식 개발 분석 절차

○ 증발량 산출 모형

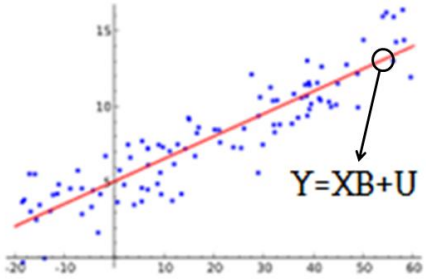
증발량 산출을 위해 선형기법 중 일반선형모형(GLM, General Linear Model)과 비선형기법, 기계학습 기반의 랜덤포레스트(RF, Random Forest) 모형을 고려하였으며 소형증발량 산출 모형 2개와 대형증발량 산출 모형 2개, 총 4개의 모형을 구축하여 PM 산출식과 비교하고자 하였다.

➢ **General Linear Model (GLM, 일반선형모형)**

- 선형 모형 상에서 하나 이상의 변수를 대상으로 일반화된 모형을 구축하는 선형 기법이다.

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

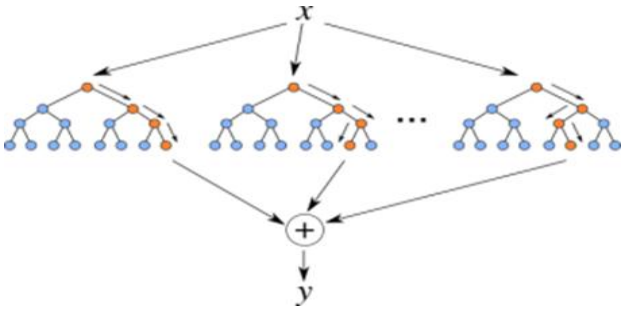
- 전통적인 선형 모형이 갖는 '종속변수의 정규분포'와 '분산의 동등성' 가정을 배제하고 자료의 독립성 가정과 모형의 가법성(additivity) 원리에 기초한 통계 모형이다. 종속변수의 분포에 따라 다양한 선형회귀기법을 사용할 수 있는 특징을 가진다.



<Linear Regresstion Model>

➢ **Random Forest (랜덤포레스트)**

- 다수개의 의사결정나무를 만들고 각각의 예측값들을 조합하여 정밀도가 높은 분류를 하는 앙상블(Ensemble)기법으로, 동일한 하나의 데이터 집합에서 임의 복원 샘플링과 학습을 여러 번 수행하여 각 예측결과를 종합하여 도출하는 비선형 기법이다.
- 모형을 구성하는 나무의 수를 ntrees, 각 나무의 최대 깊이를 max depth라고 한다.
- 복잡한 비선형 상관관계와 상호작용 효과 표현이 가능하고, 다른 분석기법에 비해 예측의 안정성과 정확도가 높은 특징을 가진다.



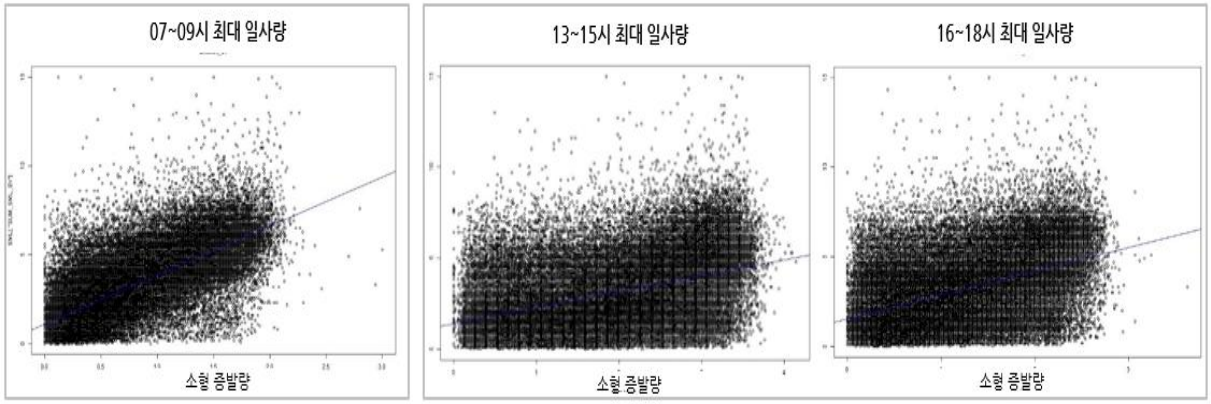
GLM 모형과 RF 모형의 최종 선택 변수는 아래 [표 2.7]와 같이 일사량, 지면온도, 상대습도, 풍속, 강수량, 눈, 일조시간, 현지기압, 평균고도, 해안면적비율 등 각각 9개이며, 이는 변수중요도를 고려하여 구축할 수 있는 모든 모형 중 예측력이 가장 우수한 모형을 최종 모형으로 선택하였을 때 구성된 변수이다. GLM 모형의 경우, 각 변수별 계수 추정과 요소별 해석이 가능하지만, RF 모형의 경우는 계수 추정이 불가능하나 변수중요도를 산출할 수 있다. 변수중요도는 각 변수의 상대적 영향을 의미한다.

[표 2.7] 증발량 산출 모형별 최종 선택 변수

소형증발량				대형증발량			
GLM		RF		GLM		RF	
변수	회귀계수	변수	변수중요도	변수	회귀계수	변수	변수중요도
07～09시 최대일사량	+0.262	07～09시 평균일사량	1	07～09시 최대일사량	+0.237	07～09시 평균일사량	1
평균 지면온도	+0.154	평균 지면온도	0.504	평균 지면온도	+0.051	평균전운량	0.400
07～09시 최저상대습도	-0.051	07～09시 최저상대습도	0.207	강수유무	-0.027	평균 지면온도	0.248
2M풍속	+0.038	평균전운량	0.154	최대풍속	+0.023	07～09시 최고상대습도	0.218
일강수량	-0.019	07～09시 최대풍속	0.112	일강수량	-0.022	일강수량	0.162
일최심적설	-0.015	16～18시 최저풍속	0.107	평균현지기압	-0.016	07～09시 최대풍속	0.161
합계일조시간	+0.010	ASOS 평균고도	0.103	평균상대습도	-0.014	16～18시 최저풍속	0.155
해안면적비율	+0.008	일강수량	0.061	19～21시 평균일사량	+0.014	ASOS 평균고도	0.118
ASOS 평균고도	+0.007	해안면적비율	0.026	해안면적비율	+0.010	해안면적비율	0.041

최종 선택된 변수들을 살펴보면 모든 모형에 '07~09시 최대일사량', '07~09시 최저상대습도', '07~09시 최대풍속' 등 07~09시의 기상변수가 선택되었음을 확인할 수 있다. 소형증발량과 07~09시의 기상변수 값의 관계, 소형증발량과 그 외 시간의 기상변수의 관계를 보기 위하여 [그림 2.6]과 같이 산점도를 그려보았다. 산점도 결과 '07~09시 최대일사량'의 경우 소형증발량과 양적 선형관계를 보이고 있으나 낮, 오후시간의 '13~15시 최대일사량', '16~18시 최대일사량'의 경우 소형증발량과 '07~09시 최대일사량'에 비해 뚜렷한 관계를 보이고 있지 않음을 확인 할 수 있다. 따라서 일반적으로 증발이 잘 되는 13~18시에 비해 07시~09시에 기상조건에 따른 증발량의 변화가 뚜렷하여 증발량 산출 모형에 더 영향력이

높은 변수로 선택되었다고 판단된다.



[그림 2.6] 시간별 최대 일사량과 소형증발량의 관계

2.1.5. 검증 결과 및 결론

앞서 구축한 모형의 성능과 연도별, 지점별 검증 결과를 살펴보고, 그에 따른 결론을 논하고자 한다.

❍ 증발량 산출 모형 성능

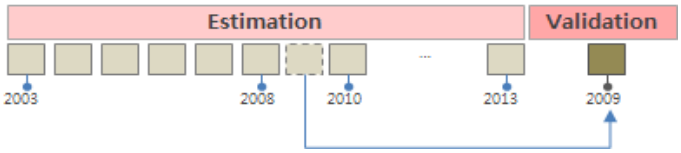
모형설명력(R^2)과 예측오차를 통하여 각 모형의 성능을 비교하면 [표 2.8]과 같다. 이 때, 모형설명력(R^2)는 구축된 모형을 통해 데이터를 설명할 수 있는 비율이라고 할 수 있으나 모형설명력이 높다고 예측을 잘하는 모형이라고 할 수는 없다. 모형 예측오차는 모형 구축 시 학습에 사용된 데이터를 활용하여 검증한 평균 절대 백분율 오차이며, 평균 예측오차는 모형 학습에 사용되지 않은 데이터를 활용하여 연도별로 검증(교차 검증)한 평균 절대 백분율 오차의 평균값이다. 소형과 대형 증발량 두 경우 모두 RF 모형의 모형설명력이 GLM 모형보다 더 높고, 예측오차가 더 작기 때문에 GLM보다 RF의 모형 성능이 더 좋다고 할 수 있다.

[표 2.8] 증발량 산출 모형의 성능

	소형증발량		대형증발량	
	GLM	RF	GLM	RF
모형설명력(R^2)	56.29%	63.27%	40.60%	47.69%
모형예측오차	18.60%	11.01%	18.96%	12.27%
평균예측오차	18.57%	17.05%	18.85%	17.70%

➢ **K-fold cross-validation (교차검증)**

- 모형 도출에 사용할 데이터 셋과 검증에 사용할 데이터 셋을 분리하여 검증기간의 분석 모형에 의한 예측값($\hat{Y}$)과 실제값(Y)의 차이로 예측모형의 정확도를 검증한다.
- k번의 예측오차를 평균하여 개발모형의 성능을 파악한다.



➢ **대칭 평균 절대 백분율 오차 (SMAPE, Symmetric Mean Absolute Percentage Error)**

- 예측에 대한 정확도를 나타내는 기준의 하나이다.
- 실제값과 산출된 적합치(또는 예측치)의 차이에 대한 비율 절대값을 평균하여백분율로 나타낸 것으로 작은 값일수록 좋은 성능임을 검증하는 기준이다.

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{|Y_i| + |\hat{Y}_i|}$$

[표 2.9] 모형별 평균예측오차 비교

	PM 산출식	GLM	RF
소형 증발량 평균예측오차	18.94%	18.57%	17.05%
대형 증발량 평균예측오차	19.43%	18.85%	17.70%

GLM 모형과 RF 모형 결과와 PM 산출식 결과를 비교하면 위 [표 2.9]와 같다. 평균예측오차 값은 소형과 대형 증발량 모두 PM 산출식이 가장 크고, 다음으로 GLM, RF순으로 작아졌다. 이 결과는 기존의 산출식보다 GLM과 RF의 모형이 조금 더 예측을 잘한다고 해석할 수 있다. 하지만 이는 지점별 교차 검증한 결과를 상세히 살펴보면 다른 결론을 내릴 수 있다.

[표 2.10]의 지점별 평균예측오차와 [표 2.11]의 PM산출식과 예측모형 간의 예측오차 차이를 고려했을 때, PM 산출식의 예측력이 더 좋은 지점은 춘천, 북강릉, 인천, 서산, 안동이고 PM 산출식의 예측력이 더 낮은 지점은 철원, 울릉도, 여수이다. 그 외 9개 지점에서는 비슷한 성능을 보였다.

하지만 PM산출식보다 좋은 결과를 보여준 세 지점(철원, 울릉도, 여수)은 일사량을 측정하지 않는 지점이다. 모형에 입력되는 일사량 값들은 근사한 지점의 값으로 대체된 데이터였으며 이를 통하여 구축된 모형이 PM 산출식보다 더 나은 성능을 보였다고 하기 어렵다. 따라서, 오랜 시간 동안 연구된 PM 산출식이 증발량 산출 모형보다 성능이 좋거나 비슷하다고 결론지을 수 있다.

[표 2.10] 증발량 산출 모형의 지점별 예측오차(%)

	소형증발량			대형증발량		
지점번호	GLM모형	RF모형	PM산출식	GLM모형	RF모형	PM산출식
95(철원)	28.03	22.30	35.71	27.99	23.73	39.46
101(춘천)	18.97	18.10	15.24	20.21	19.62	15.61
104(북강릉)	23.55	18.97	12.14	23.57	21.16	16.45
105(강릉)	20.69	18.36	21.87	21.78	20.49	20.04
108(서울)	17.92	17.80	17.30	19.02	18.63	17.34
112(인천)	19.18	18.42	14.68	20.24	19.71	16.24
115(울릉도)	20.73	20.35	34.74	20.80	20.16	39.10
119(수원)	16.84	16.14	13.49	16.99	16.46	13.32
129(서산)	17.04	16.43	13.80	16.30	15.62	11.77
131(청주)	17.18	15.69	15.76	16.60	15.87	13.70
133(대전)	18.31	16.17	20.60	16.73	15.67	15.67
136(안동)	17.14	15.99	12.33	17.63	16.71	13.62
146(전주)	18.63	17.62	19.55	17.13	16.77	15.45
159(부산)	15.04	13.00	17.13	15.73	13.26	15.64
165(목포)	16.92	15.40	14.12	16.83	15.91	12.61
168(여수)	16.21	14.87	29.80	19.04	17.23	40.50
184(제주)	16.98	16.17	15.32	18.56	17.71	15.91
평균예측오차	**18.79**	**17.16**	**19.03**	**19.13**	**17.92**	**19.56**

[표 2.11] PM산출식과 증발량 산출 모형의 오차평균 차이

	PM산출식	GLM모형		RF모형		PM vs RF	
지점번호	오차평균	오차평균	차이	오차평균	차이		
95(철원)	**37.59**	**28.01**	**9.57**	**23.20**	**14.38**	**3**	**RF**
101(춘천)	15.42	19.59	-4.17	18.85	-3.43	15	PM
104(북강릉)	14.29	23.56	-9.26	19.98	-5.69	17	PM
105(강릉)	20.95	21.23	-0.28	19.54	1.42	6	비슷
108(서울)	17.32	18.47	-1.15	18.21	-0.89	8	비슷
112(인천)	15.46	19.71	-4.25	19.09	-3.63	16	PM
115(울릉도)	**36.92**	**20.76**	**16.16**	**20.2**	**16.66**	**2**	**RF**
119(수원)	13.41	16.92	-3.51	16.32	-2.91	12	비슷
129(서산)	12.78	16.67	-3.89	16.02	-3.24	13	PM
131(청주)	14.73	16.89	-2.16	15.80	-1.07	9	비슷
133(대전)	18.14	17.52	0.61	15.87	2.27	5	비슷
136(안동)	12.98	17.39	-4.41	16.36	-3.39	14	PM
146(전주)	17.50	17.88	-0.38	17.15	0.35	7	비슷
159(부산)	16.38	15.39	1.00	13.16	3.23	4	비슷
165(목포)	13.37	16.87	-3.51	15.67	-2.30	11	비슷
168(여수)	**35.15**	**17.62**	**17.52**	**16.06**	**19.09**	**1**	**RF**
184(제주)	15.62	17.77	-2.15	16.93	-1.31	10	비슷

※ PM vs RF : PM 오차평균과 RM 오차평균의 차이 값을 기준으로 성능 우위 판별

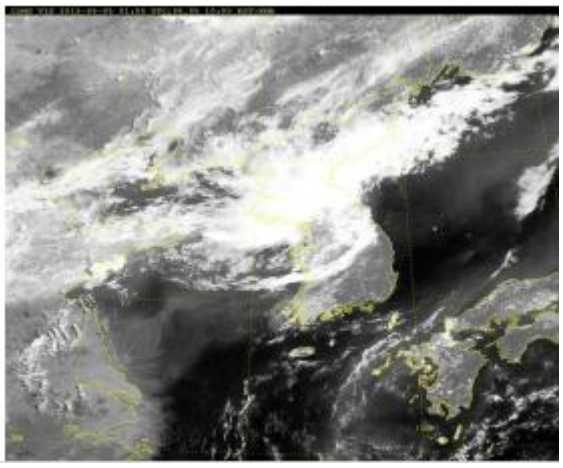
청 내 빅데이터 분석 서비스

(관측) 운량 산출식 개발

2.2. 운량 산출식 개발

2.2.1. 개요

운량은 항해, 항공, 농업 및 야외활동 등, 일상생활에 깊은 영향을 미치는 기상 요소 중의 하나이다. 대기에서 운량은 전 지구 기후시스템을 조절하는 태양에너지의 강도 조절의 역할과 함께, 지구로부터 방출되는 장파복사[6]가 우주 밖으로 나가는 정도를 조절하는 역할을 한다.



COMS 가시영상

운량산출은 일반적으로 관측자에 의한 목측을 통해 생산되어 활용되어 왔다. 목측을 통한 운량 산출의 경우 하늘전체(눈으로 보이는 범위)의 몇 %가 구름으로 덮여있는지를 판단하고, 이를 0부터 10까지의 계급화하여 산출한다. 구름의 두께나 종류에는 관계없이 하늘전체의 30%가 덮여있으면 3으로 한다. 하지만 개별 관측자에 의한 주관적인 영향 및 일기상태에 따라 정확한 운량의 산출이 어려우며, 해상에서의 관측 및 연속적인 관측이 불가능하고, 지점에서 관측한 값이 넓은 영역을 대표하는 데에는 한계가 있다는 단점이 존재한다. 따라서 넓은 영역에 대한 고품질의 객관적 운량산출정보가 필요한 시점이다.

현재 유인 기상관서에서 매시 운량 및 운형을 관측하고 있으며, 관측자동화를 위해 도입한 운고·운량계에서도 운량, 운형자료가 측정되고 있다. 또한 천리안 위성 등에서 운량 및 운형 정보가 산출되고 있다. 이 자료간의 관계를 분석하여 도출된 관계식을 활용할 수 있게 된다면 운고·운량계 정확도를 검증하고 무인관측 지점에도 구름 관측 자료를 산출할 수 있게 되는 등 관측 업무 자동화에 큰 기여를 할 것으로 기대되어 본 분석을 시행하게 되었다.

2016년 청 내 빅데이터 분석 서비스 중 운량 분석 과제는 목측기반, 운고계 및 위성 기반의 운량 간의 관계를 빅데이터 기법을 통하여 관계식을 도출하는 것으로 정의하며, 천리안 위성 및 운고운량계 분석과 히마와리 위성 및 관측 정보 분석 2가지를 시행하였다. 본 기술노트에 천리안 위성 및 운고운량계 분석, 히마와리 위성 및 관측 정보 분석 순으로 기술되었고, 운량 산출식 개발 분석 과제에 대한 요약은 [표 3.1]과 같다.

---

6) 장파복사(長波輻射, long wave radiation) : 지구에서 내보내는 복사를 말하며, 파장이 짧은 태양복사에 대해 지구복사의 파장이 긴 것을 이르는 말이다. 지구복사는 1~100μm의 파장에 적외선 영역에 속한다.

[표 3.1] 운량 산출 분석별 분석 과제 요약

		천리안위성 및 운고운량계 분석	히마와리 위성 및 관측 정보 분석
데이터	공통	목측 운량, ASOS 기상관측 정보	
	기상	운고운량계 운량, 천리안위성 운량	히마와리 위성 운량, 히마와리 위성 채널간 관측 정보 차이값
	비기상	위·경도, 산림면적, 도시화율, 해발고도평균, 해면부, 일출·일몰시간	-
분석 기간		2013.01 ~ 2015.07 (2년 7개월)	2016.05 ~ 2016.08 (4개월)
분석 대상		28개 관측소	22개 관측소
선정 기준		목측, 천리안위성, 운고운량계에서 운량에 대한 공통 관측정보가 존재하는 기간과 관측소	지속적으로 운량을 관측하고 있는 22개 관측소
분석 기법		다항 로지스틱 회귀분석	랜덤 포레스트(Random Forest)
오차 범위별 정확도	±0	50.7%	58.6%
	±1	64.6%	73.3%
	±2	78.6%	86.2%

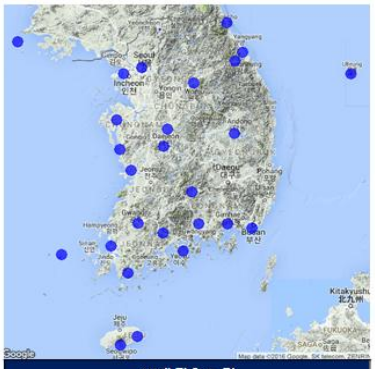
2.2.2. 천리안 위성 및 운고운량계 분석

❍ 데이터 수집 및 분석 대상 선정

천리안 위성 및 운고운량계 분석을 통한 운량 산출식 개발을 위해 수집된 데이터 목록은 [표 3.2]와 같다. 기상 데이터는 ASOS의 목측 구름 자료와 기온, 강수량, 풍속, 일조량, 기압, 안개, 증발량 등과 운고·운량계 데이터, COMS(천리안위성)의 천리안위성 전운량, 운정고도 데이터를 활용하였다. 비기상 데이터로는 위도, 경도, 도시화율, 해발고도평균, 해면부면적 등을 활용하였다.

목측 관측소 기준으로 2013~2015년간 1회 이상 관측 정보가 존재하는 99개소 중 관측 정보가 60% 이상 존재하는 40개소 후보를 먼저 선정하였다. 그 중 목측, 천리안위성, 운고운량계에서 운량에 대한 공통 관측 정보가 존재하는 2013년 1월부터 2015년 7월까지에 해당하는 전국 28개 관측소 최종 선정하였다.

이 때, 구름관측이 유지되고 있는 관측소(22개소) 중 7개소(춘천, 수원, 포항, 울산, 제주, 전주, 대구)는 관측자료의 수집 불안정으로 분석 대상에서 제외되었다.



28개 관측소 명			
90 속초	114 원주	155 창원	172 고창
95 철원	115 울릉도	156 광주	174 순천
100 대관령	129 서산	159 부산	188 성산
102 백령	131 청주	165 목포	189 서귀포
104 북강릉	133 대전	168 여수	192 진주
108 서울	136 안동	169 흑산도	235 보령
112 인천	140 군산	170 완도	284 거창

[표 3.2] 천리안 위성 및 운고운량계 분석 수집 데이터

<table>
<tr><th>기상기후 데이터</th><th>비(非)기상 데이터</th></tr>
<tr><td>ASOS</td><td>지형데이터</td></tr>
<tr><td>수집기간 :
(목측구름) 2011.01~2016.08 (5년8개월, 1시간 단위)
(기상관측) 2010.01~2015.12 (6년)
지점 관측소 : (목측구름) 99, (기상관측) 102
데이터건수 : (목측구름) 4,621,000건
(기상관측) 5,363,568건
변수종류 :
(목측구름) 목측 전운량, 중하층운량, 중하층운고, 운형
(기상관측) 기온, 강수량, 풍속, 습도, 일조량, 기압, 안개, 증발량</td><td>수집기간 : 2014 (단, 산림면적은 2010, 1년 단위)
지점 관측소 : 28
데이터건수 : 28건
변수종류 : 위도, 경도, 도시화율, 해발고도평균, 해면부면적
출처 : 통계청, 산림청, 기상청</td></tr>
<tr><td>운고운량계</td><td>일출/일몰시간 데이터</td></tr>
<tr><td>수집기간 : 2011.10~2016.07 (5년 10개월, 1분 단위)
지점 관측소 : 100
데이터건수 : 131,265,974건
변수종류 : 운고운량계 운량, 하층운고, 충증운고, 상층운고</td><td rowspan="3">수집기간 : 2013.01. ~ 2015.12. (1일 단위)
지점 관측소 : 36
데이터건수 : 39,420건
변수종류 : 일출 시간, 일몰 시간
출처 : KASI(한국천문연구원)</td></tr>
<tr><td>COMS (천리안 위성)</td></tr>
<tr><td>수집기간 : 2013.01~2015.07 (2년 7개월, 15분 단위)
지점 관측소 : 95
데이터건수 : 8,415,496건
변수종류 : 천리안위성 전운량, 운정고도</td></tr>
</table>

❍ 데이터 탐색 및 결측치·이상치 데이터 처리

데이터 탐색은 데이터 현황 및 특성을 파악하고, 데이터 전처리 과정을 통해 분석 목적에 적합한 데이터 셋을 정의하기 위한 선행과정이다.

각 데이터별 결측치 탐색 결과 [표 3.3]와 같은 결측율을 확인했으며 결측 데이터는 분석에서 제외되었다.

❍ 데이터 전처리 기준

목측 운량 데이터는 전운량과 중하층운량, 중하층운고 간의 로직체크[7]를 통해 데이터를 전처리하였으며 천리안위성의 운량데이터는 전운량과 운정고도, 운고운량계 데이터는 운량과 하층운고 간의 로직체크를 통하여 전처리 하였다. 자세한 기준은 부록 2-1, 2-2, 2-3에 첨부되어 있다.

7) 로직체크 : 논리적으로 타당한 결과인지 검토하는 과정을 말한다.

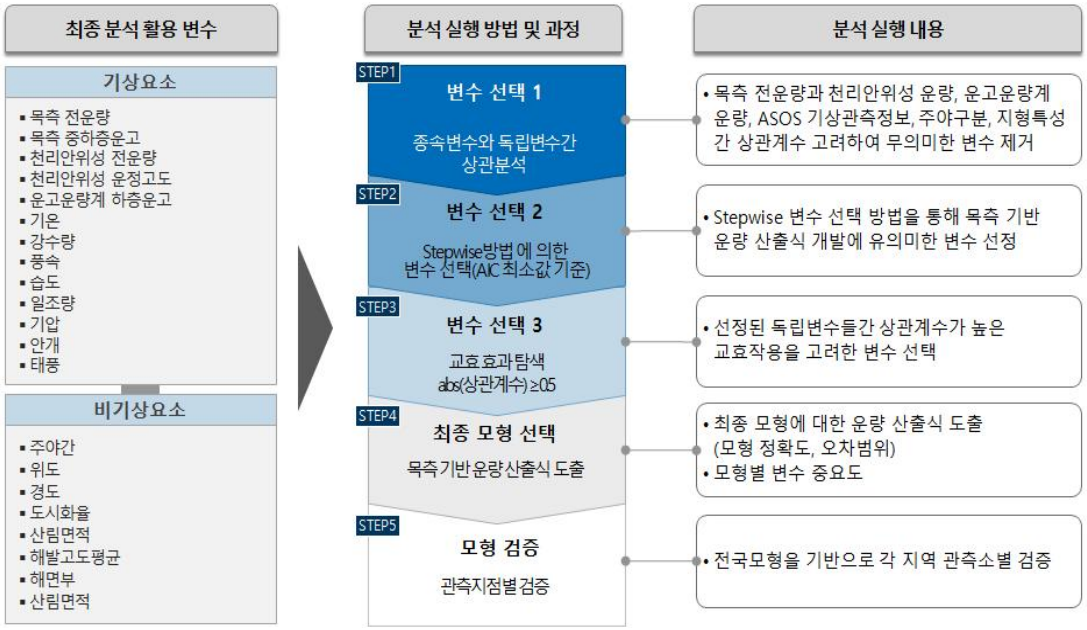
[표 3.3] 데이터별 결측치 탐색 결과

		총 관측행	결측(NA)행	결측율	비고
목측	전운량	474,768	28,318	5.96%	결측 데이터는 분석에서 제외
천리안위성	전운량	474,768	616	0.13%	
운고운량계	운량	474,768	38,186	8.0%	전체 결측 데이터의 약 27%는 고창 관측소의 관측 기기장애 → 최종분석 대상에서 제외 그 외 결측데이터 분석에서 제외
기타 기상요소	기온	474,768	57	0.01%	
	습도	474,768	62	0.01%	
	풍속	474,768	94	0.02%	
	강수량	474,768	425,519	89.63%	비가 오지 않는 날의 경우 미관측 값 → 0으로 대체
	일조량	474,768	115,819	24.39%	21~04시 일조량 미관측 시간 → 0값 대체
	기압	474,768	39	0.01%	
	안개	474,768	413	0.09%	
	증발량	474,768	229,004	48.2%	분석대상 관측지점의 커버리지 46%로 최종변수에서 제외
비기상 데이터	도시화율	474,768	0	0.00%	
	해발고도평균	474,768	0	0.00%	
	해면부	474,768	0	0.00%	
	산림면적	474,768	0	0.00%	
	위도	474,768	0	0.00%	
	경도	474,768	0	0.00%	

❍ 분석 실행 및 결과

- 분석 실행 절차

천리안 위성 및 운고운량계 분석을 통한 운량 산출식 개발을 위하여 최적의 변수 선택 후 분석 모형을 구축하고 최종 모형을 기반으로 적합도 검증을 수행하였다.



[그림 3.1] 천리안 위성 및 운고운량계 분석 절차

- 파생변수 생성 및 변수 선택

분석 정확성 향상과 유의미한 정보를 찾기 위해 원시데이터로부터 여러 변수의 조합, 조정을 통해 파생변수를 생성하였다. 온도, 강수량, 풍속, 습도, 일조량, 기압의 1시간, 3시간 전 관측값과, 천리안 위성 전운량, 운정고도의 1시간 전 관측값, 운고운량계 운량, 하층운고의 1시간 전 관측값을 파생변수로 추가하고 변수는 3단계를 거쳐 선택되었다. 먼저, 종속변수와 독립변수간의 상관분석을 통해 목측 전운량과 상관관계가 낮은 지형 변수를 삭제하고 Stepwise 방법[8]으로 유의미한 변수만 선택하였다. 마지막으로 독립변수간 교호작용[9] 효과를 고려한 3개의 변수를 최종 변수에 포함하였다.

[표 3.4] 최종 분석 활용 변수

구분	변수명	변수 설명
원천 변수	목측 전운량	목측으로 관측되는 전체 운량
	천리안위성 전운량	천리안위성으로 관측되는 전체 운량
	운고운량계 운량	AWS에서 관측되는 하늘 일부분 운량
	강수량	강수량
	습도	습도
	일조량	태양광선이 실제 비친 시간
	기압	대기 전체압력 중 그 대기에 함유된 수증기가 갖고 있는 분압
	주야구분	일출시간과 일몰시간을 기준으로 낮과 밤 구분
파생 변수	천리안위성 전운량 b1	1시간 전 천리안위성 전체 운량
	운고운량계 운량 b1	1시간 전 시점의 운고운량계 운량
	습도 b1	1시간 전 시점의 습도
교호 작용	천리안위성 전운량 * 천리안위성 전운량 b1	천리안 전운량과 1시간 전 천리안위성 전운량과의 교호효과
	운고운량계 운량 * 운고운량계 운량 b1	운고운량계 운량과 1시간 전 운고운량계 운량과의 교호효과
	일조량*습도	일조량과 습도간의 교호효과

- 분석 모형 정의

종속변수가 0~10 사이의 값을 가지는 범주형/순서형 데이터인 특성을 고려하여 다항 로지스틱 회귀모형과 순서형 로지스틱 회귀모형을 고려하였으나 더 나은 성능으로 목측 운량의 유효값인 0~10의 모든 수준에 대해 산출되는 다항 로지스틱 회귀모형을 최종 선택하였다.

8) Stepwise (단계적) 변수 선택법 : 기준 통계치에 도움이 되지 않는 변수를 삭제하거나 모형에서 빠져있는 변수 중 기준 통계치를 개선시키는 변수를 추가하는 과정을 반복하여 적절한 변수를 선택한다.
9) 교호 작용 : 독립변수 사이에 상호작용이 있어서 두 효과의 합이 산술적으로 예상되는 결과가 나타나지 않는 상태

데이터 종류	데이터 특성	적용 가능 분석 기법
운량 데이터	0 ~ 10 사이의 값을 가지는 범주형 데이터	다항 로지스틱 회귀모형
	0 ~ 10 사이의 값을 가지는 순서형 데이터	순서형 로지스틱 회귀모형

운량 산출 모형 정의

<다항 로지스틱 회귀> 선택

모형식: $\ln\left(\frac{p(\text{목측.운량}_j)}{p(\text{목측.운량}_0)}\right) = b_{0j} + b_{1j}X_1 + \cdots + b_{kj}X_k, j = 1,\ldots,10$

✓ 목측 운량 수준에 대해 다항분포를 가정

➔ 장점: 운량 산출값이 0~10의 모든 수준에 분포함

<순서 로지스틱 회귀>

모형식: $\ln\left(\frac{p(\text{목측.운량} \le j)}{1-p(\text{목측.운량} \le j)}\right) = b_{0j} + b_1X_1 + \cdots + b_kX_k, j = 1,\ldots,10$

✓ 목측 운량 수준에 대해 순서형을 가정

➔ 단점: 운량 산출값이 0~10의 모든 수준에 분포하지 않고, 산출값이 주로 운량이 많은 경우에만 존재하여 분석 모형에서 배제

목측운량 ~ f(상수항, 천리안위성 전운량, 운고운량계 운량, ASOS, 지형특성, 파생변수, 교호작용)

[그림 3.2] 분석 모형 정의

- 분석 결과 : 최종 분석 모형 및 회귀계수

13개 독립변수를 이용한 10개의 다항 로지스틱 회귀 모형을 통하여 목측 전운량을 산출한다. 최종 모형은 AIC[10]를 기준으로 적합한 모형을 선택하였다. 최종 모형의 회귀계수는 아래 [그림 3.3]과 같다. 모형1은 운량이 1일 때의 확률을 산출하며 운량이 0인 경우 1-(모형1+모형2+…+모형10)의 확률값으로 계산된다.

다항 로지스틱 회귀 모형 : $\ln\left(\frac{p(\text{목측.운량}_j)}{p(\text{목측.운량}_0)}\right) = b_{0j} + b_{1j}X_1 + \cdots + b_{kj}X_k, j = 1,\ldots,10$

추정된 회귀 계수

모형 번호	(Intercept)	천리안위성 전운량	운고운량계 운량	강수량	습도	일조량	기압	주야 구분	천리안위성 전운량.b1	운고운량계 운량.b1	습도.b1	천리안위성 전운량: 천리안위성 전운량.b1	운고운량계 운량: 운고운량계 운량.b1	습도:일조량
1	-4.4924	0.1762	0.3993	0.6383	0.0075	-0.1292	0.0376	1.8281	0.0978	0.3983	-0.0013	-0.0143	-0.0361	-0.0064
2	-4.4022	0.4549	0.3514	0.5335	0.0260	-0.6894	0.0107	1.6984	0.3853	0.2509	-0.0233	-0.0610	-0.0223	-0.0092
3	-4.2750	0.5069	0.3023	0.6699	0.0391	-0.6999	0.0229	1.1287	0.3105	0.3485	-0.0342	-0.0638	-0.0305	0.0028
4	-5.1788	0.5945	0.3576	0.6815	0.0727	-0.9410	-0.0042	1.4427	0.3953	0.3335	-0.0607	-0.0779	-0.0193	0.0025
5	-6.1919	0.6393	0.4270	0.5330	-0.0298	-0.7380	0.0497	1.7068	0.3106	0.3437	0.0344	-0.0595	-0.0282	-0.0005
6	-5.6052	0.6585	0.4638	0.3159	0.0500	-1.4712	0.0280	1.1458	0.3342	0.3253	-0.0425	-0.0671	-0.0366	0.0155
7	-5.8524	0.4972	0.4274	-3.0044	0.0044	-2.4037	0.0590	1.6338	0.1685	0.4398	-0.0043	-0.0274	-0.0383	0.0252
8	-4.4733	0.4374	0.4822	-0.1749	0.0315	-3.1454	0.0445	2.0249	0.1255	0.3604	-0.0327	-0.0234	-0.0399	0.0231
9	-5.8902	0.3856	0.5580	0.6029	-0.0092	-3.1327	0.0562	2.6194	0.0222	0.4736	0.0016	-0.0043	-0.0442	0.0167
10	-8.6269	0.0835	0.5824	0.4920	0.0794	0.5868	0.0501	2.4141	-0.3518	0.5088	-0.0416	0.0664	-0.0448	-0.0539

[그림 3.3] 분석 모형 및 추정 회귀계수

10) AIC : Akaike Information Criterion, 모형 적합도를 의미하며 최종 모형 선택의 기준으로 활용된다. AIC 값이 상대적으로 작을수록 좋은 모형임을 나타낸다.

- 모형 검증 결과

최종 분석 모형을 통해 산출된 운량 산출의 정확도는 아래 [그림 3.4]를 통해 확인할 수 있다. ±0 오차범위 50.7%, ±1 오차범위 64.6%, ±2 오차범위 78.6% 수준으로 확인되었다. 운량 0이나 10의 경우 예측 정확도가 매우 높은 편이지만, 그 외 예측 정확도는 상대적으로 낮은 수준이다. 이는 운량 데이터의 분포가 0과 10에 집중되어 있기 때문으로 추정된다.

산출 운량 →

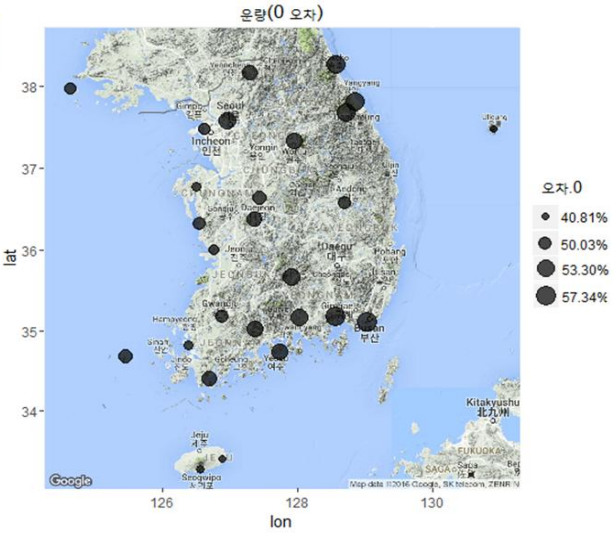
목측 운량 ↓	0	1	2	3	4	5	6	7	8	9	10	±0오차범위	±1오차범위	±2오차범위
0	93.6%	1.0%	0.3%	0.6%	1.4%	0.5%	0.7%	0.1%	0.8%	0.0%	0.9%	93.6%	94.6%	94.9%
1	88.8%	1.7%	0.8%	2.1%	2.0%	1.6%	1.1%	0.7%	0.7%	0.0%	0.3%	1.7%	91.3%	93.4%
2	76.5%	2.6%	0.9%	3.6%	4.4%	3.7%	3.4%	1.1%	2.7%	0.1%	1.0%	0.9%	7.0%	87.9%
3	60.7%	2.6%	1.0%	4.3%	7.3%	6.9%	5.8%	2.7%	5.7%	0.5%	2.6%	4.3%	12.6%	22.0%
4	42.3%	2.6%	0.9%	3.4%	8.8%	11.0%	8.8%	5.2%	10.3%	1.5%	5.3%	8.8%	23.2%	50.2%
5	29.1%	1.8%	0.7%	2.2%	9.4%	12.9%	9.2%	7.2%	15.0%	2.9%	9.5%	12.9%	31.5%	32.9%
6	18.4%	1.1%	0.4%	1.9%	8.4%	11.6%	7.3%	10.0%	18.6%	5.0%	17.5%	7.3%	28.8%	55.8%
7	9.0%	0.5%	0.2%	0.7%	6.5%	8.4%	4.7%	8.6%	20.4%	8.0%	33.2%	8.6%	33.6%	50.0%
8	4.0%	0.2%	0.1%	0.2%	4.8%	4.5%	2.4%	5.5%	16.9%	10.4%	51.1%	16.9%	32.7%	86.2%
9	1.1%	0.1%	0.0%	0.1%	3.1%	2.3%	0.9%	3.3%	12.3%	11.0%	65.9%	11.0%	89.2%	92.4%
10	0.4%	0.0%	0.0%	0.0%	2.8%	0.7%	0.4%	0.5%	3.8%	3.6%	87.7%	87.7%	91.3%	95.1%
											정확도	50.7%	64.6%	78.6%

[그림 3.4] 운량 산출 모형의 예측 적합도

각 지역관측소별 운량 산출 모형 검증을 수행하였다. 27개 지역 관측소별 정확도는 ±0 오차범위 기준 최대 57.3%, 최저 40.8% 수준이며, ±1 오차범위 기준 최대 70.6%, 최소 56.9%, ±2 오차범위 기준 최대 82.3%, 최소 74.3%임을 확인하였다.

〈 관측소별 결과 검증 〉

관측소(STN_ID)	±0 오차범위	±1 오차범위	±2 오차범위
창원(155)	57.3%	70.6%	82.3%
부산(159)	56.7%	70.0%	81.2%
속초(90)	56.6%	69.8%	81.4%
대관령(100)	54.3%	67.1%	79.6%
북강릉(104)	53.6%	69.1%	81.3%
거창(284)	53.6%	65.1%	78.1%
진주(192)	53.4%	67.6%	79.4%
서울(108)	53.3%	69.0%	81.5%
순천(174)	53.2%	64.7%	78.1%
여수(168)	52.8%	64.4%	77.3%
원주(114)	51.6%	65.8%	79.4%
대전(133)	51.5%	66.0%	80.2%
완도(170)	50.8%	63.4%	76.6%
철원(95)	50.7%	65.1%	78.9%
청주(131)	50.6%	66.3%	79.4%
흑산도(169)	50.4%	59.8%	75.5%
안동(136)	50.4%	64.0%	77.4%
광주(156)	50.0%	64.9%	79.7%
보령(235)	49.9%	64.9%	78.8%
백령(102)	49.7%	64.8%	79.0%
인천(112)	48.5%	62.8%	75.8%
군산(140)	48.1%	61.7%	77.1%
목포(165)	47.8%	60.5%	76.5%
서산(129)	46.2%	60.2%	76.7%
서귀포(189)	43.4%	57.1%	74.3%
성산(188)	41.9%	56.9%	76.7%
울릉도(115)	40.8%	59.9%	78.5%



✓ 상기 그림은 ±0 오차 정확도에 기초한 결과
→ 원이 클 수록 정확도가 높은 관측소

[그림 3.5] 관측소별 운량 산출 모형 검증

2.2.3. 히마와리 위성 및 관측 정보 분석

❍ 데이터 수집 및 분석 대상 선정

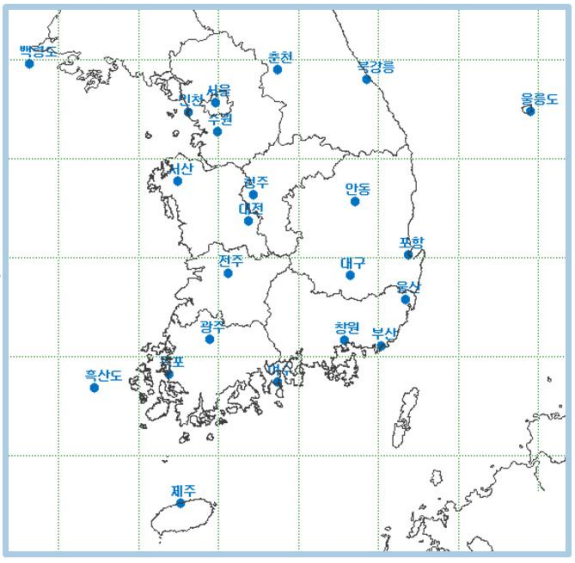
히마와리 위성 및 관측 정보 분석을 통한 운량 산출식 개발을 위해 수집된 데이터 목록은 [표 3.5]와 같다.

[표 3.5] 수집 데이터와 분석 대상

<table>
<tr><th>ASOS</th><th>히마와리 위성</th></tr>
<tr><td>수집기간 :
(목측구름) 2016.05~2016.08 (4개월, 1시간 단위)
(기상관측) 2016.05~2016.08 (4개월)
지점 관측소 : (목측구름) 22, (기상관측) 22
데이터건수 : (목측구름) 64,042건
(기상관측) 64,042건
변수종류 :
(목측구름) 목측 전운량, 중하층운량, 운형
(기상관측) 기온, 초상온도, 지면온도</td><td rowspan="3">수집기간 : 2016.05~2016.08 (4개월, 1시간 단위)
지점 관측소 : 22
데이터건수 : 64,042건
변수종류 : 히마와리위성 운량,
대상격자영역 내 격자점수(G1),
구름판단 격자점수(G2),
위성 채널간 차이값 평균

가시광선 (반사도)
적외선 (온도값)

천리안1 | 히마와리 (채널, 파장(um), 해상도(km)) | 비고
1 | 0.46 | 1 | 가시(B)
2 | 0.51 | 1 | 가시(G)
VIS(0.68) | 3 | 0.64 | 0.5 | 가시(R)
4 | 0.86 | 1
5 | 1.6 | 2
6 | 2.3 | 2
IR4(3.7) | 7 | 3.9 | 2
8 | 6.2 | 2 | 수증기(450hPa)
IR3(6.8) | 9 | 7.0 | 2 | 수증기(500hPa)
10 | 7.3 | 2 | 수증기(750hPa)
11 | 8.6 | 2 | SO2
12 | 9.6 | 2 | O3
IR1(10.8) | 13 | 10.4 | 2 | 대기창
14 | 11.2 | 2 | 대기창
IR2(12.0) | 15 | 12.3 | 2 | 대기창
16 | 13.3 | 2 | CO2

히마와리 위성을 이용한 운량 산출 분석 대상 채널
1) 관측소 지점에서16.4Km 반경 내 13번 채널을 기준으로 위성값-지면값의 차이가 6.5 이상이면 구름격자로 처리
2) 위 반경 내 전체 영역(G1), 구름 없는 영역(G1-G2), 구름 영역(G2)에 대한 채널간 차이값의 평균을 계산 (채널 7~16번 대상)</td></tr>
<tr><th>분석 대상 관측소</th></tr>
<tr><td>지속적으로 운량을 관측하고 있는 22개 관측지점
 
</td></tr>
</table>

히마와리 위성 총 16개 채널 중 적외선에 의한 위성값 및 지면값 온도 측정이 가능한 7~16번 채널을 대상으로 운량 산출식 개발을 위한 채널간 차이값 평균 데이터를 산출하여 분석에 활용하였다.

본 분석을 위해 수집된 데이터의 변수와 해당 변수의 속성을 정리하면 [표 3.6]과 같다.

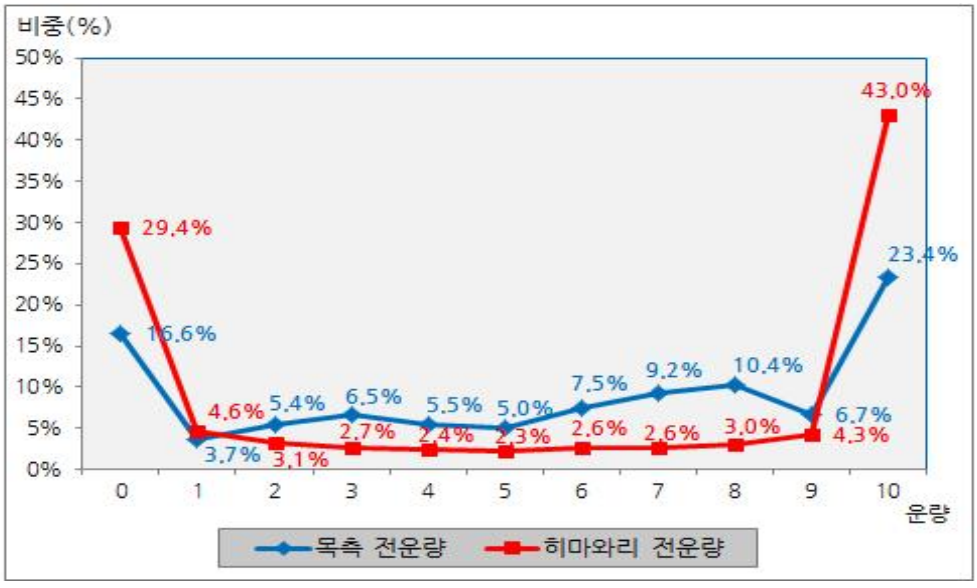
[표 3.6] 수집 데이터의 변수별 속성

변수		속성	단위
목측 운량	목측 전운량	목측 전운량	0～10
	목측 중하층 운량	목측 중하층 구름의 운저운고	100미터
	목측 운형	목측 운형	-
기상관측 정보	기온	사람 키 높이 기온	℃
	초상온도	풀 위의 온도	℃
	지면온도	땅위의 온도	℃
히마와리 위성	대상 격자영역내 격자점수(G1)	대상 격자영역내 격자점수	격자점수
	구름으로 판단되는 격자점수(G2)	구름으로 판단되는 격자점수	격자점수
	채널7-8: 차이값 평균(G1)	G1영역에서 7채널과 8채널 차이값의 평균	
	채널7-9: 차이값 평균(G1)	G1영역에서 7채널과 9채널 차이값의 평균	
	(이하 생략) → 채널 7～16까지 서로 다른 두 개의 채널 차이 값 채널간 차이값 평균 변수 : Ch7-8, Ch7-9, … Ch7-16, Ch8-9, Ch8-10, … Ch9-Ch10, … Ch15-16 등 (총 45개)		

❍ 데이터 탐색 및 결측치·이상치 데이터 처리

데이터 탐색은 데이터 현황 및 특성을 파악하고, 데이터 전처리 과정을 통해 분석 목적에 적합한 데이터 셋을 정의하기 위한 선행과정이다. 목측 운량, 히마와리 위성 데이터, ASOS 기상관측 정보 등으로 구성된 분석 데이터를 대상으로 결측치 및 이상치 파악, 데이터 분포 확인 등의 데이터 탐색을 수행하였다.

목측 운량과 히마와리 위성 데이터의 전운량 분포는 [그림 3.6]과 같이 0과 10에 집중되어 있는 형태이다. 또한 결측치 탐색 결과는 [표 3.7]과 같다.



[그림 3.6] 전운량 분포

[표 3.7] 데이터별 결측치 및 이상치 탐색 결과

		총 관측행	결측(NA)행	결측율	비고
목측 운량	전운량	64,042	13,876	21.7%	결측 및 이상치 데이터는 분석에서 제외
	중하층 운량	64,042	10,342	6.2%	
	운형	64,042	23,017	35.9%	운형의 NA는 “구름없음” 의미 운량과의 로직 체크를 통해 분석 대상 여부 정의
ASOS 기상요소	기온	64,042	142	0.02%	결측 또는 미관측 데이터의 비율이 낮은 상태로 전처리 과정을 통해 변환 또는 제외
	초상온도	64,042	130	0.02%	
	지면온도	64,042	130	0.02%	
히마와리 위성 데이터	전운량	64,042	0	0.00%	운량 산식 = round((G2격자점수/G1격자점수)*10, 0) G1 : 대상 격자 영역 내 격자점수 G2 : 구름으로 판단되는 격자점수
	11-16채널 차이 평균(G1)	64,042	0	0.00%	모든 채널차이평균 변수 중 유의미한 몇 가지만 예시로 표에 포함
	13-16채널 차이 평균(G1)	64,042	0	0.00%	
	15-16채널 차이 평균(G1)	64,042	0	0.00%	
	7-13채널 차이 평균(G1)	64,042	2	0.00%	채널간 차이값이 계산불가인 경우 분석 대상 제외

❍ 데이터 전처리 기준

분석에 활용될 목측 운량 데이터는 전운량과 중하층운량, 중하층운고 사이의 논리적 타당성을 고려하여 전처리 하였으며 그 기준은 부록2-4에 첨부하였다.

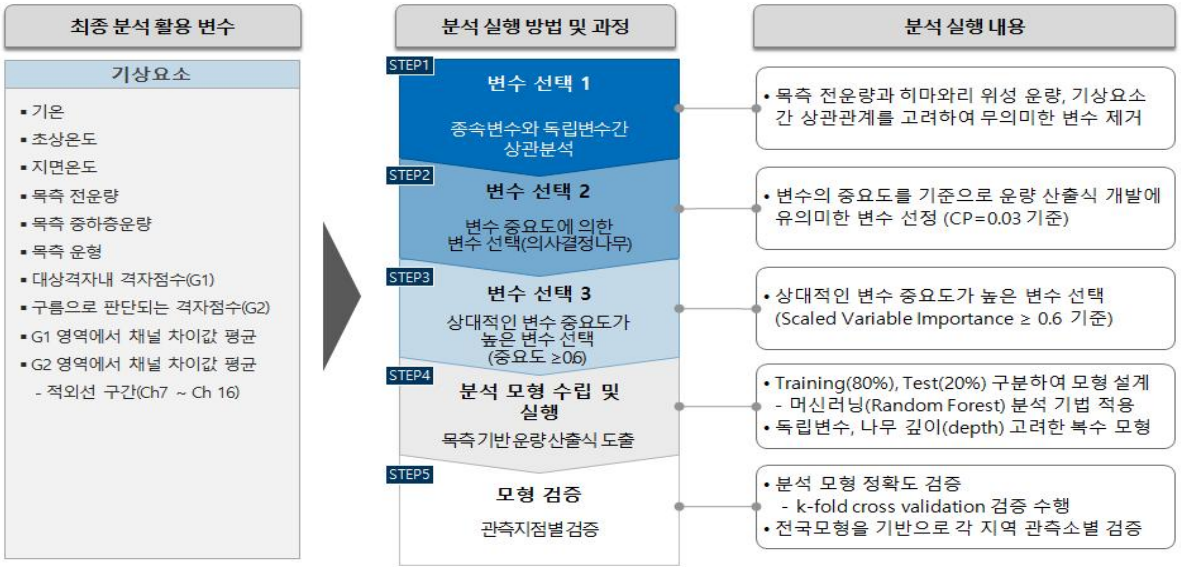
목측 전운량 데이터가 결측(NA)인 경우, 음수(-9)인 경우는 결측 데이터로 분석에서 제외하였으며, 목측 전운량이 0보다 크고 운형이 없는 경우는 안개에 의한 운량으로 데이터 분석에서 제외하였다. 목측 전운량이 0이지만 운형이 있는 경우 또한 데이터 분석에서 제외하였으며 기상 요소 중 -99.9인 경우도 데이터 분석에서 제외하는 전처리를 하였다.

❍ 분석 실행 및 결과

- 분석 실행 절차

히마와리 위성 및 관측 정보 분석을 통한 운량 산출 모형을 구축하기 위하여 먼저 변수 선택 과정을 진행하였다. 정제된 50개의 변수 중에서 상관분석과 의사결정나무 모형을 통해 변수 중요도 기준으로 14개의 변수를 선정하였다.

14개의 변수로 이루어진 최종 분석 데이터를 학습용 데이터(80%), 검증용 데이터(20%)로 구분하고 훈련용 데이터로 랜덤포레스트 모형을 구축하였다. 랜덤포레스트 모형에서 산출되는 변수 중요도를 고려하여 최종 8개의 변수를 선택하고 최종 모형을 선정하였다. 마지막으로 검증용 데이터로 분리되었던 20%의 데이터로 모형의 성능을 확인하고 교차검증을 통해 분석 모형에 대한 검증을 수행하다.



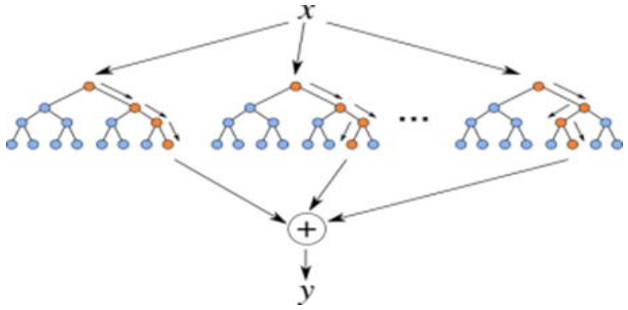
[그림 3.7] 히마와리 위성 및 관측 정보 분석 절차

- 분석 기법 및 최종 모형

기계학습 기법 중 하나인 랜덤포레스트를 이용하여 운량 산출 모형을 구축하였으며 이 기법은 과적합[11]을 방지하고 여러 모델의 예측 결과를 종합하여 정확도를 높이는 기법 중 하나이다. 모형에 입력되는 독립변수의 수, 나무의 깊이(Depth)를 조정하며 여러 모형을 구축한 뒤 비교적 적은 변수로 우수한 성능을 가지는 모형을 최종 모형을 선정하였다. 그 결과는 [표 3.8]과 같다.

➢ **Random Forest (랜덤포레스트)**

- 다수개의 의사결정나무를 만들고 각각의 예측값들을 조합하여 정밀도가 높은 분류를 하는 앙상블(Ensemble)기법으로, 동일한 하나의 데이터 집합에서 임의 복원 샘플링과 학습을 여러 번 수행하여 각 예측결과를 종합하여 도출하는 비선형 기법이다.
- 모형을 구성하는 나무의 수를 ntrees, 각 나무의 최대 깊이를 max depth라고 한다.
- 복잡한 비선형 상관관계와 상호작용 효과 표현이 가능하고, 다른 분석기법에 비해 예측의 안정성과 정확도가 높은 특징을 가진다.



11) 과적합 : over fitting, 학습데이터를 과하게 잘 학습하여 추후 발생하는 데이터에 대한 예측력이 좋지 않은 현상

[표 3.8] 최종 랜덤포레스트 모형

최종 모형			모형 성능 평가		
			적합도	검증	
독립변수 수	8	accuracy	모형(80%) (39,268건)	검증(20%) (9,817건)	교차 검증 (49,085건)
트리수(ntree)	128	± 0 오차범위	58.6%	46.2%	45.5%
Max(depth)	10	± 1 오차범위	73.7%	66.5%	66.3%
결정계수(R^2)	0.9603	± 2 오차범위	86.2%	82.0%	82.1%
※ 최종 변수 G1영역에서 Ch11과 Ch16 차이값 평균, G1영역에서 Ch10과 Ch11 차이값 평균, G1영역에서 Ch13과 Ch16 차이값 평균, G1영역에서 Ch12와 Ch14 차이값 평균, G1영역에서 Ch07과 Ch13 차이값 평균, G1영역에서 Ch07과 Ch14 차이값 평균, G1영역에서 Ch15와 Ch16 차이값 평균, G1영역에서 Ch08과 Ch11 차이값 평균					

2.2.4. 검증 결과 및 결론

최종 모형의 성능은 [그림 3.8]을 통하여 확인할 수 있다. 랜덤포레스트 모형을 통한 운량 산출의 정확도는 ±0 오차범위 58.6%, ±1 오차범위 73.7%, ±2 오차범위 86.2% 수준이다.

산출 운량 →

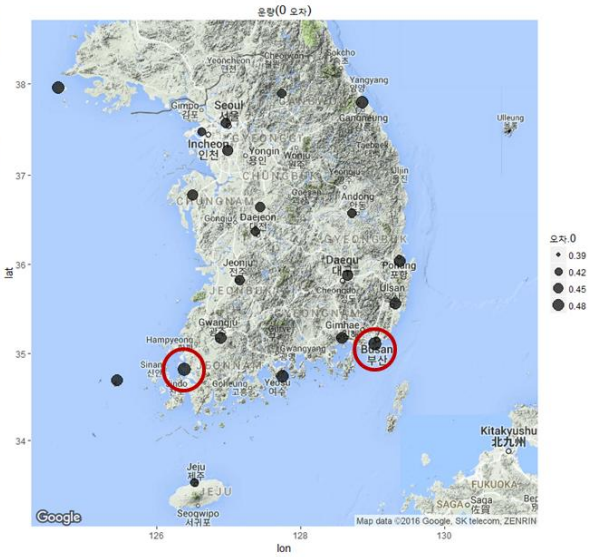
목측운량 ↓	0	1	2	3	4	5	6	7	8	9	10	±0오차범위	±1오차범위	±2오차범위
0	97.1%	0.1%	0.1%	0.9%	0.3%	0.2%	0.6%	0.4%	0.1%	0.0%	0.1%	97.1%	97.2%	97.3%
1	71.4%	19.1%	1.1%	5.1%	1.7%	0.6%	0.9%	0.1%	0.1%	0.0%	0.0%	19.1%	91.6%	96.7%
2	58.8%	1.2%	20.6%	9.8%	2.2%	1.4%	3.8%	1.2%	0.5%	0.0%	0.5%	20.6%	31.7%	92.6%
3	38.8%	0.6%	1.1%	39.9%	3.3%	2.3%	9.0%	3.3%	0.8%	0.0%	0.9%	39.9%	44.3%	47.1%
4	24.9%	0.4%	1.8%	11.4%	30.1%	2.9%	16.4%	8.6%	1.6%	0.1%	1.9%	30.1%	44.4%	62.6%
5	13.6%	0.2%	0.9%	10.7%	3.8%	28.6%	19.4%	14.7%	3.7%	0.1%	4.2%	28.6%	51.8%	62.6%
6	9.5%	0.2%	0.6%	5.5%	1.7%	1.1%	45.4%	18.6%	8.0%	0.7%	8.9%	45.4%	65.0%	74.7%
7	5.9%	0.1%	0.2%	2.5%	1.2%	0.5%	9.1%	48.5%	12.3%	1.0%	18.8%	48.5%	69.9%	71.4%
8	2.7%	0.0%	0.2%	1.3%	0.5%	0.1%	5.4%	10.8%	38.8%	1.7%	38.6%	38.8%	51.3%	95.2%
9	1.5%	0.0%	0.1%	0.8%	0.4%	0.0%	2.1%	6.6%	12.8%	19.2%	56.4%	19.2%	88.4%	95.0%
10	0.6%	0.0%	0.0%	0.2%	0.0%	0.0%	0.6%	2.2%	2.8%	0.4%	93.1%	93.1%	93.5%	96.3%
											정확도	58.6%	73.7%	86.2%

[그림 3.8] 운량 산출 모형의 예측 적합도

도출된 모형 운량 산출식을 기반으로 각 지역관측소별 운량 산출 모형을 적용하여 검증을 수행하였다. 지역 관측소별 ±0 오차범위 최대는 50.5%(부산)이며, 최소는 38.7%(울릉도), ±1 오차범위 기준 최대는 71.9%(북강릉), 최소 61.0%(울릉도)로 나타났다. ±2 오차범위 기준 최대는 86.4%(북강릉), 최소 75.1%(울릉도)로 나타나 울릉도의 운량 예측의 정확성이 가장 낮고 부산, 북강릉 지점의 운량 예측 정확성이 높다는 것을 [그림 3.9]를 통하여 확인할 수 있다.

〈 관측소별 결과 검증 〉

관측소(STN_ID)	±0 오차범위	±1 오차범위	±2 오차범위
부산(159)	50.5%	69.6%	83.4%
목포(165)	50.1%	66.8%	80.6%
백령(102)	49.0%	66.1%	81.3%
북강릉(104)	48.5%	71.9%	86.4%
여수(168)	48.5%	65.4%	82.4%
포항(138)	48.0%	69.5%	85.1%
흑산도(169)	47.7%	60.2%	78.7%
광주(156)	47.5%	69.2%	82.8%
창원(155)	47.3%	68.6%	84.6%
울산(152)	47.2%	68.8%	84.9%
대구(143)	46.4%	69.1%	85.0%
수원(119)	46.2%	68.6%	84.8%
서산(129)	45.4%	67.3%	83.8%
청주(131)	44.9%	68.2%	84.4%
서울(108)	44.9%	71.2%	86.2%
전주(146)	44.8%	65.7%	81.2%
대전(133)	44.1%	67.8%	84.5%
춘천(101)	44.1%	67.4%	84.6%
안동(136)	43.7%	65.5%	82.9%
인천(112)	42.4%	65.8%	83.3%
제주(184)	42.2%	66.5%	80.8%
울릉도(115)	38.7%	61.0%	75.1%



✓ 상기 그림은 ±0 오차 정확도에 기초한 결과

→ 원이 클 수록 정확도가 높은 관측소

[그림 3.9] 관측소별 운량 산출식 개발 모형 검증

앞서 히마와리 위성 자료와 관측 정보를 바탕으로 구축한 랜덤포레스트 모형은 2016년 5월부터 8월까지의 데이터로 구축하고 2016년 9월 데이터로 검증하였기에 계절적 특징을 반영하지 못하여 예측력이 기대보다 낮은 수준이었다고 판단된다. 또한, 본 분석에 활용된 히마와리 8호 위성은 일본에서 2014년 발사하고 2015년 가동을 시작하여 데이터가 안정적으로 생산되기 시작한 기간이 길지 않아 짧은 기간을 분석할 수밖에 없었다.

운량 데이터의 분포가 0과 10에 밀집되어 있다는 근본적인 원인으로 특히 운량 1, 2, 5, 9와 같은 경우 예측 정확도가 낮다. 하지만 차원 축소 혹은 군집분석, 모형의 parameter, cut-off value 조정 등을 통하여 모형을 고도화 한다면 더 나은 결과를 도출할 수 있을 것으로 기대된다.

이번 청 내 빅데이터 분석 서비스를 통하여 위성 자료로 운량을 산출할 수 있다는 가능성이 검증된 것에 의의를 가지며 더 많은 데이터가 구축된 뒤 고도화된 모형으로 운량 관측 자동화가 이뤄지기를 바래본다.

청 내 빅데이터 분석 서비스

(항공) 항공기 출발 지연 분석

2.3. 항공운항 영향 기상 변수 임계치 도출

2.3.1. 개요

지난 2016년 1월 23일 7년 만에 제주도에 발효된 한파주의보와 대설특보, 강풍특보 등으로 제주국제공항의 항공기 운항에 차질을 빚는 등 관광객과 도민이 큰 불편을 겪는 사건이 있었다.

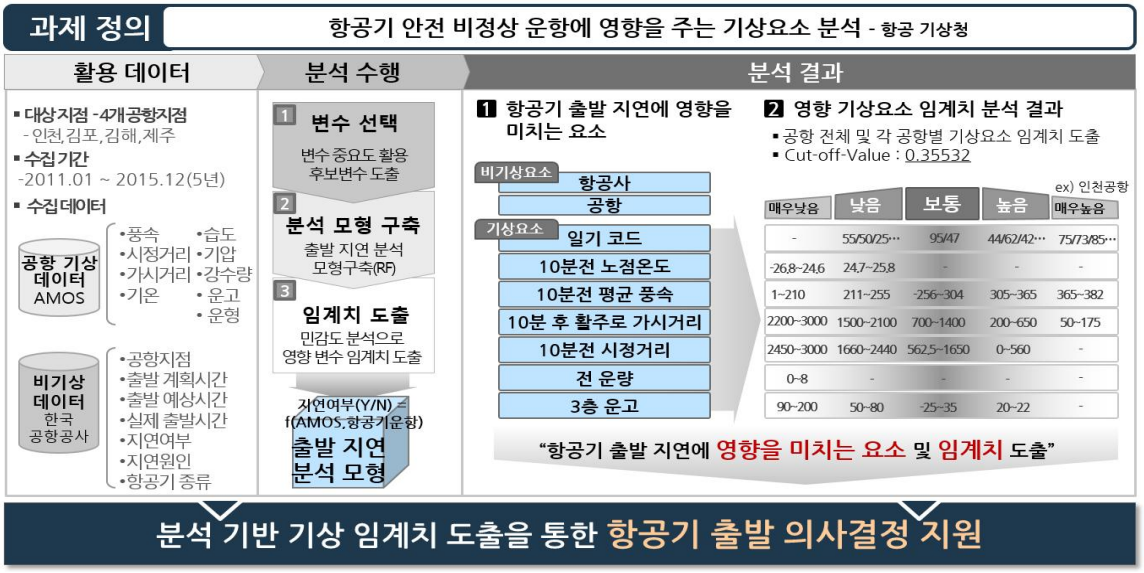
출발·도착편 117편이 결항했고, 73편이 지연 운항했다. 이로 인해 제주공항 터미널에 대기 승객 4천여 명의 발이 묶였다.

△제주공항 결항 지연 속출<출처=연합뉴스>

이처럼 날씨는 항공운항에 큰 영향을 주는 요인이기 때문에 태풍, 안개, 폭설 등 위험 기상이 항공기 비정상운항(결항, 지연)에 미치는 영향을 분석하고, 이를 바탕으로 항공기 운항에 영향을 미치는 국내 공항별 취약성을 고려한 영향예보 제공과 의사결정 지원이 필요하다. 2017년 6월 목표로 국토교통부에서 구축 중인 항공교통종합통제센터(ATFMC)에서는 항공로에 대한 정량적이고 체계적인 흐름관리 업무 수행을 위해 위험기상(뇌전(TS), 바람 등)과 항공교통 흐름관리 상관도를 고려한 새로운 기상정보의 필요성이 대두되고 있다.

이에 기상청에서는 공항별 출발·도착지별 기상에 따른 비정상운항 원인과 위험기상의 상관성을 분석하여 항공기 지연/결항에 의한 정상운항 감소율과 기상요소별 영향 임계치를 도출하고자 한다. 시정, 운고 등 기상요소가 공항 및 공역(항공로 상)의 항공기 비정상 운항에 미치는 영향도 도출이 가능하다면 항공운항에 필요한 영향예보 제공의 기반이 마련될 것으로 기대되어 본 분석을 시행하게 되었다.

2016년 청 내 빅데이터 분석 서비스 중 항공운항 과제는 원활한 항공교통 흐름관리를 위해 항공기 출발 지연에 영향을 미치는 기상요소를 분석하고, 기상요소별 항공기 출발 지연에 영향을 미치는 임계치를 도출하는 것으로 정의한다. 항공기 출발지연에 영향을 미치는 기상요소 분석 과제에 대한 요약은 다음 [그림 4.1]과 같다.



[그림 4.1] 항공기 출발 지연 분석 과제 요약

2.3.2. 데이터 수집 및 탐색

❍ 데이터 수집 및 분석 대상 선정

항공운항 영향 기상변수 분석을 위해 공항기상관측 데이터 확보가 가능한 4개 공항지점(인천, 김포, 제주, 김해)을 분석 대상으로 선정하고 2011년부터 2015년까지 5년 동안의 항공기 운항 데이터, AMOS(공항기상관측) 데이터를 수집하였다.

[표 4.1] 분석 대상 및 수집 데이터

(기상기후) AMOS 공항기상관측
수집기간 : 2011.01 ~ 2015.12 (5년) 지점 관측소 : 4개 관측소의 대표활주로 지점 - 인천(113), 김포(110), 김해(153), 제주(182) 데이터건수 : (1분 단위) 10,136,502 건 (1시간 단위) 175,293 건 변수종류 : 풍향, 풍속, 시정, 운량, 운고, 운형, 기온, 습도, 이슬점온도, 현지기압, 해면기압, 강수량 등
(비기상) 공항 운항 데이터
수집기간 : 2011.01 ~ 2015.12 (5년) 지점 관측소 : 4개 관측소의 대표활주로 지점 - 인천(113), 김포(110), 김해(153), 제주(182) 데이터건수 : 1,545,595 건 변수종류 : 항공기 편명, 공항지점, 출발(도착)계획시간, 출발(도착)예상시간, 실제 출발시간 , 지연여부, 지연원인, 결항여부, 결항원인, 결항원인 공항, 항공기 종류 출처 : KAC 한국 공항 공사

수집된 데이터의 변수는 [표 4.2]와 같다. 공항운항 데이터의 중 인천공항은 출발 계획시간 데이터가 없어 출발 예상시각변수와 같은 값으로 설정하였다.

[표 4.2] 수집 데이터 내 변수

데이터	변수		데이터	변수	
공항 운항 데이터	날짜	항공사	AMOS (1시간)	1층 운량	3층 운량
	출발 공항	편명		1층 운형	3층 운형
	도착 공항	항공기 종류		1층 운고	3층 운고
	출발 계획 시간	지연 여부		2층 운량	4층 운량
	출발 예상 시간	지연 원인		2층 운형	4층 운형
	실제 출발 시간			2층 운고	4층 운고
				4677코드일기현상	전운량
AMOS (1분)	관측시각	시정거리 1분	AMOS (1분)	활주로 가시거리 1분	풍향 2분 평균
	지점번호	시정거리 1분MID		활주로 가시거리1분MID	풍향 2분 최대
	활주로방향	시정거리 10분		활주로 가시거리 10분	풍향 10분 평균
	현지기압(QFE)[12]	시정거리 10분MID		활주로 가시거리 10분MID	풍향 10분 최대
	고도계 수정치(QNH)[13]	온도		풍속 2분 평균	풍속 10분 평균
	운고1층	노점온도		풍속 2분 최대	풍속 10분 최대
	습도	강수량1분			

❍ 데이터 탐색 및 결측치·이상치 데이터 처리

데이터 탐색은 데이터 현황 및 특성을 파악하고, 데이터 전처리 과정을 통해 분석 목적에 적합한 데이터 셋을 정의하기 위한 선행과정이다.

- 기상요인으로 인한 비정상 운항 정의

공항 운항 데이터 중 정상운항은 92.6%, 지연운항 7.3%, 결항은 0.7%로 탐색되었다. 여러 가지 원인으로 발생된 지연운항 중 기상이나 기타 원인으로 인한 지연운항을 기상요인으로 인한 비정상 운항 데이터로 정의하고 분석 데이터에 포함하였다.

▪ 정상 운항 및 비정상 운항 탐색

공항명	데이터 수 (건)	정상운항		지연운항		결항	
		수(건)	비율	수(건)	비율	수(건)	비율
인천공항	676,810	634,008	93.7%	41,011	6.1%	1,791	0.3%
김해공항	192,319	184,337	95.8%	7,982	4.2%	1,541	0.8%
김포공항	339,596	317,706	93.6%	21,890	6.4%	3,846	1.1%
제주공항	336,870	294,914	87.5%	41,956	12.5%	3,396	1.0%
총데이터수	1,545,595	1,430,965	92.6%	112,839	7.3%	10,574	0.7%

▪ 지연원인이 기상 및 기타인 경우

공항명	지연운항(건)	지연 원인(원인공항이 자신일 경우)	
		기상 원인	기타원인
인천공항	41,011	656(1.6%)	17,091(41.7%)
김해공항	7,982	170(2.1%)	171(2.1%)
김포공항	21,890	489(2.2%)	159(0.7%)
제주공항	41,956	237(0.6%)	378(0.9%)
총 데이터 수	112,839	1,552(1.4%)	17,799(15.8%)

[그림 4.2] 비정상 운항 정의를 위한 운항 데이터 탐색

12) QFE : 현지기압, field elevation pressure. 공항공식표고에서의 기압값
13) QNH : 고도계 수정치, atmospheric pressure at nautical height. 현지기압에 ICAO 표준대기 값을 적용하여 해면 경정한 값

- 결측치·이상치 데이터 처리

기상 데이터의 이상치는 항공기후통계지침(부록 3-1 참고)을 기준으로 판별하였으며 운항 데이터와 기상 데이터의 결측치는 [그림 4.3]과 같은 기준으로 처리하였다.

공항 운항 데이터

- **실제 출발시간(ATT)이 결측인 경우,**
  - 결측 데이터 분석에서 제거
- **예상 출발시간(ETT)이 결측인 경우,**
  계획 출발시간(STT)으로 대체

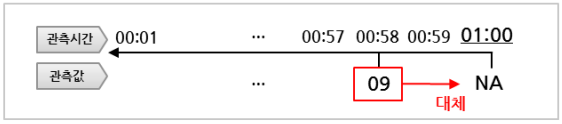
시간 단위 AMOS 기상 데이터

- 4677코드일기현상, 운량, 운형 변수에서 0은 의미있는 값
  - NA 값은 -999로 치환
- 전운량 > 0 인 경우, 1층 운량,운형,운고가 결측이면 Knn-Imputation 기법을 활용하여 결측 대체
- 전운량 = 0 인 경우, NA를 -999로 대체

분 단위 AMOS 기상 데이터

- **결측치 데이터 시간 기준으로 이전 60분 이내에서 가까운 시간의 관측값으로 대체**

ex) 관측시간 "01:00"의 결측값을 대체할 경우

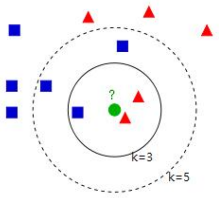


- 이전 60분 이내에 관측값이 하나도 없는 경우, -999로 대체하여 결측값으로 유지 → 모형 학습 시 학습 데이터에서 제외
- **1분단위 기상자료 중 연속적으로 1시간 이상 결측인 경우**
  - 데이터 분석 제외

[그림 4.3] 결측치 및 이상치 처리 기준

➢ kNN(k Nearest Neighbor)-Imputation

◆ 데이터 결측치 존재 시, 예측하고자 하는 데이터로부터 가장 가까운 k개의 이웃을 찾은 뒤 이들 이웃으로부터 예측하고자하는 데이터의 분류를 정하여 값을 대체하는 방법



2.3.3. 데이터 전처리

❍ 파생변수 생성

분석 정확성 향상과 유의미한 정보를 찾기 위해 원시데이터로부터 여러 변수의 조합, 조정을 통해 의미 있는 파생변수를 생성하는 과정이다.

본 분석에서는 AMOS 분 단위 데이터를 활용하여 기준 시간 이전 10분 동안의 기상상태 및 이후 10분 동안의 기상상태를 파생변수로 생성하였다. 풍속, 풍향의 경우 AMOS 데이터에 이미 10분간 기상 변수 데이터가 존재하여 파생변수 생성 대상에서 제외하였으며 운고1층, 시정거리, 가시거리와 같이 시야방해와 관련된 변수에 대해서는 기준 시간 전·후 10분 동안의 평균값, 최댓값, 최솟값 변수를 생성하였다. 노점온도, 습도, 기압과 같은 변수는 이전 10분 동안에 대하여 파생변수를 생성하였다. 이 때, 노점온도와 운고1층의 경우 결측치 대체 전 데이터를 활용한 이전 10분간 기상파생변수 총 6개를 추가로 생성하였다.

	노점온도	습도	현지기압	QNH	운고1층	시정거리	가시거리	시정거리 MID	가시거리 MID
평균값	이전 10분간 기상변수 평균값				이전 10분간 기상변수 평균값 이후 10분간 기상변수 평균값				
최댓값	이전 10분간 기상변수 최댓값				이전 10분간 기상변수 최댓값 이후 10분간 기상변수 최댓값				
최솟값	이전 10분간 기상변수 최솟값				이전 10분간 기상변수 최솟값 이후 10분간 기상변수 최솟값				

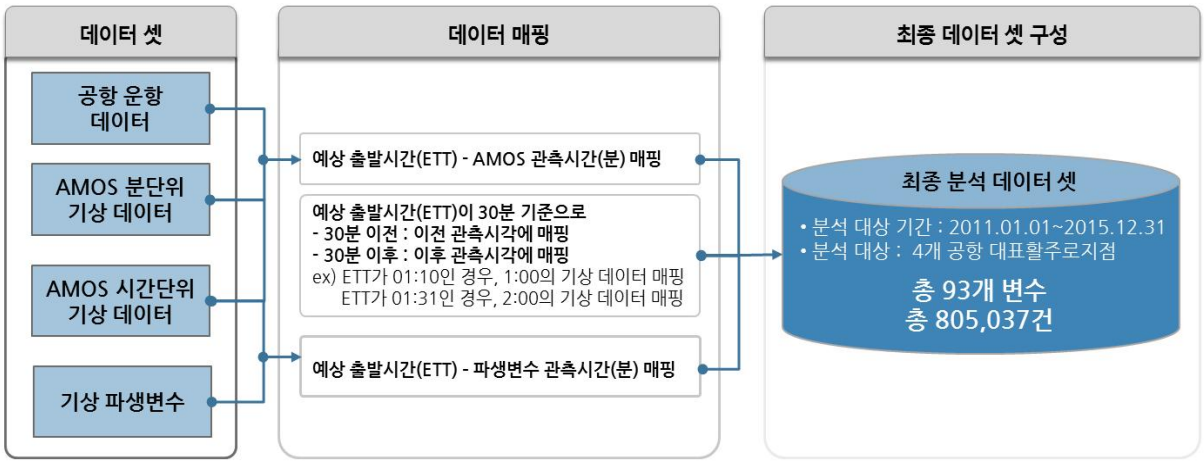
총 48*개 파생변수 생성

*노점온도 및 운고1층 변수에 대해서는, 결측 대체 전 데이터를 활용하여 이전 10분간 기상변수 평균값, 최댓값, 최솟값을 추가로 생성 (총 6개 추가 생성)

[그림 4.4] 기상파생변수 생성

❍ 데이터 매핑

분석에 활용되는 데이터들을 시간을 기준으로 결합하였다. 이 때 시간 단위와 분 단위 기상 데이터의 결합을 위하여 30분을 기준으로 매핑하였다. 1시 10분과 같이 30분 전은 이전 관측시간인 1시 데이터와 결합하고 1시 31분과 같이 30분 이후 데이터는 이후 관측시간인 2시 데이터와 결합하였다.



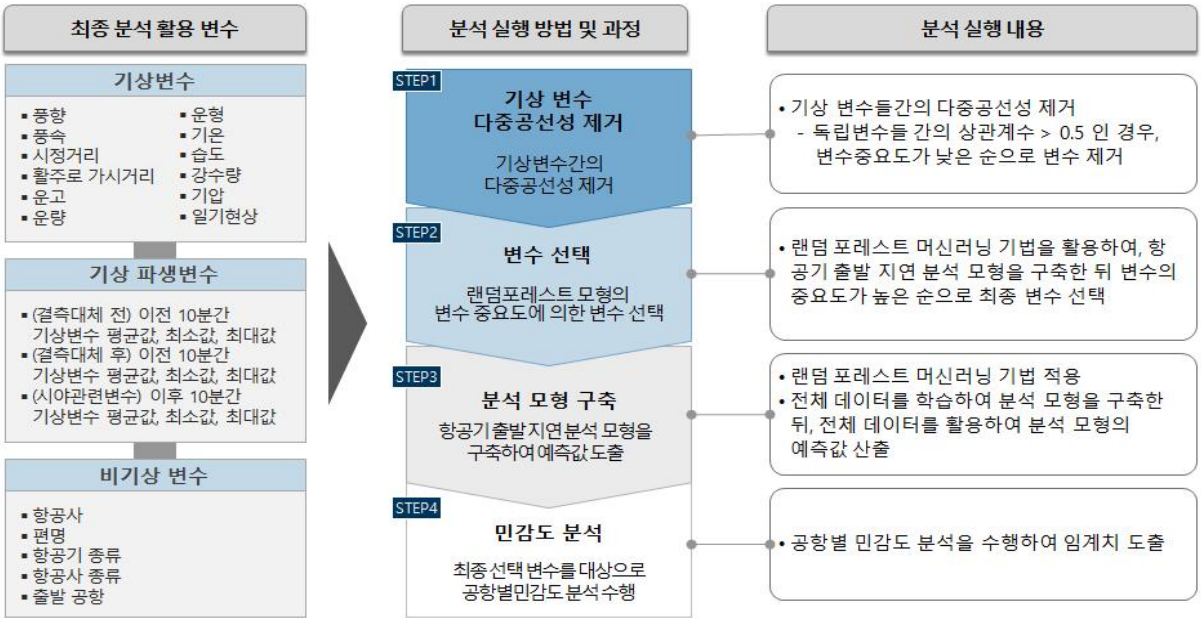
[그림 4.5] 항공기 출발 지연 분석 데이터 셋

2.3.4. 분석 실행 및 결과

❍ 분석 실행 절차

최종 분석 데이터 셋에서 상관계수를 기준으로 기상변수 간의 다중공선성[14]을 제거하고, 랜덤포레스트 기법을 활용하여 변수 중요도에 따른 최종 변수 선택과 항공기 출발지연 예측 모형을 구축하였다. 구축된 예측 모형을 통하여 지연 확률을 예측하고 민감도 분석을 통해 기상요소별 항공기 출발지연 위험의 임계값을 도출하였다. 자세한 분석 절차는 다음 [그림 4.6]과 같다.

14) 다중공선성 : multicollinearity 독립변수간의 상관계수가 높아 분석에 부정적인 영향을 주는 현상



[그림 4.6] 항공기 출발 지연 분석 절차

❍ 변수 선택과 분석 기법

항공기 출발 지연에 영향을 미치는 기상변수들간의 다중공선성을 제거하기 위하여 독립변수들 사이의 상관계수를 산출하고 상관계수가 0.5 이상인 경우 중요도가 낮은 변수를 제거하였다. 랜덤포레스트 기법을 이용하여 변수중요도가 높은 순으로 10개의 변수를 선택하였다.

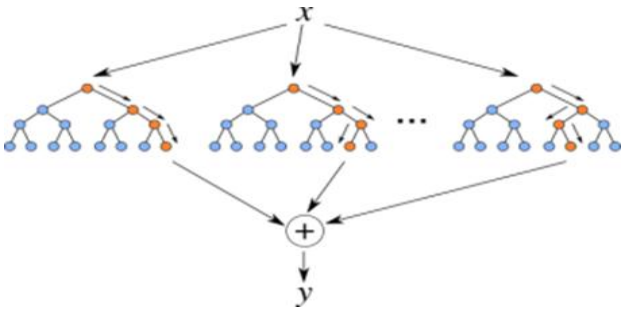
[표 4.3] 최종 변수 선택

변수	변수 중요도	비고
항공사	1	공항 운항 데이터
일기코드	0.61322	
공항	0.543758	
10분전 노점 온도	0.474718	공항 기상 관측 (AMOS)
10분전 평균 풍속	0.40833	
10분 후 활주로 가시거리	0.262659	
10분전 시정거리	0.261961	
전운량	0.166168	
3층 운고	0.132745	
4층 운고	0.007484	

항공기 출발 지연 예측 모형은 기계학습 기법 중 하나인 랜덤포레스트 기법을 활용하였으며 이 기법은 과적합을 방지하고 여러 모델의 예측 결과를 종합하여 정확도를 높이는 기법 중 하나이다.

➢ **Random Forest (랜덤포레스트)**

- 다수개의 의사결정나무를 만들고 각각의 예측값들을 조합하여 정밀도가 높은 분류를 하는 앙상블(Ensemble)기법으로, 동일한 하나의 데이터 집합에서 임의 복원 샘플링과 학습을 여러 번 수행하여 각 예측결과를 종합하여 도출하는 비선형 기법이다.
- 모형을 구성하는 나무의 수를 ntrees, 각 나무의 최대 깊이를 max depth라고 한다.
- 복잡한 비선형 상관관계와 상호작용 효과 표현이 가능하고, 다른 분석기법에 비해 예측의 안정성과 정확도가 높은 특징을 가진다.



❍ 항공기 지연확률 예측모형 성능

랜덤포레스트 기법으로 구축된 항공기 출발 지연 예측 모형의 성능을 다양한 기준을 통하여 확인하였다. 정확도(ACC)는 97.96%로 높은 수준이며 임계성공지수(CSI)는 34.99%이다. LogLoss는 0.0842로 0에 가까워 좋은 결과라고 할 수 있다. AUC는 0.9, Gini 약 0.95로 두 수치 모두 1에 가까워 본 모형이 적합한 모형임을 의미한다.

[표 4.4] 항공기 출발 지연 예측 모형 결과

구분		운항 실측		Error	모형 적합 결과	
		정상(0)	지연(1)			
운항 예측	정상(0)	ⓐ 747,025	ⓑ 4,933	0.006560207	ACC	97.96 %
	지연(1)	ⓒ 10,794	ⓓ 8,463	0.560523446	CSI	34.99 %
					LogLoss	0.084
Totals		757,819	13,396	0.020392498	AUC	0.950
					Gini	0.900

➢ **분석 모형 평가 기준**

- **정확도(Accuracy)** : $\frac{a+b}{a+b+c+d} \times 100(\%)$
- **임계성공지수(CSI)** : $\frac{d}{b+c+d} \times 100(\%)$
- LogLoss : 유무(본 분석의 경우 지연여부)에 대한 예측 검정치로 낮은 수치일수록 좋다.
- AUC : Area under the curve, 분류분석 시 민감도와 특이도의 관계를 보여주는 ROC Curve의 곡선 아래 면적을 의미하며 모형적합도의 기준이 된다. 1.0에 가까울수록 좋다.
- Gini : 민감도에 관한 예측 검정치로 1에 가까울수록 정확한 모형임을 의미한다.

2.3.5. 항공기 출발 지연 위험도 임계치 산출

❍ 항공기 출발 지연 확률과 임계치

앞서 구축된 항공기 출발지연 예측모형을 통하여 지연 발생 여부와 그에 대한 확률을 얻을 수 있다. 모형을 통해 출발 지연 확률에 대한 Cut-off value[15]는 0.355로 산출되었다.

위험도 등급별 임계치를 도출하기 위해 Cut-off value 이하를 위험도가 매우 낮은 것으로 정의하고 확률 0.15 간격으로 위험도 낮음, 보통, 높음, 매우 높음으로 정의하였다. 예측모형을 통해 출발 지연에 영향 주는 변수로 선정된 10개의 변수 중 표준화된 변수중요도(Scaled importance)가 0.01 이상인 기상변수들 즉, 4층 운고를 제외한 7개 기상 관련 변수에 대하여 출발 지연 확률과 위험도 등급별 임계치를 도출하였다.

➢ **출발 지연 확률과 위험도 등급**

- 매우 낮음 : 확률이 0.355 이하
  다른 조건들이 '낮음' 이상이 아닐 때 항공기 출발 지연 가능성이 매우 낮은 경우
- 낮음 : 확률이 0.355 초과 0.505 이하
  다른 조건들이 '낮음' 이상이 아닐 때 항공기 출발 지연 가능성이 비교적 낮은 경우로 구두 언급 등 소극적인 조치가 필요
- 보통 : 확률이 0.505 초과 0.655 이하
  다른 조건들이 '낮음' 이상이 아닐 때 항공기 출발 지연 가능성이 보통인 경우로 구두 경고 등 약간의 조치가 필요
- 높음 : 확률이 0.655 초과 0.805 미만
  다른 조건들에 관계없이 항공기 출발 지연 가능성이 높아 안내멘트 등 일반적인 조치가 필요
- 매우 높음 : 확률이 0.805 초과
  다른 조건들에 관계없이 항공기 출발 지연 가능성이 매우 높아 출발시간 변경 등 적극적인 조치가 필요

❍ 영향 요인의 공항별 항공기 출발 지연 위험도 임계치

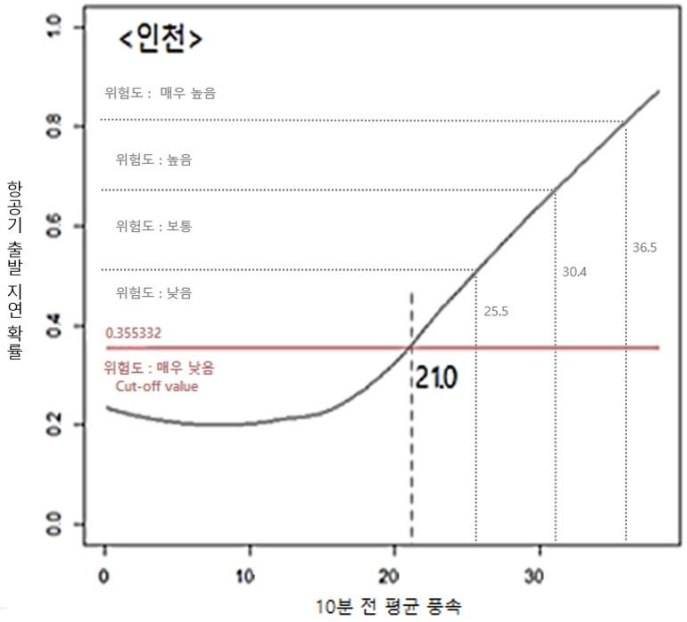
[표 4.5]는 앞서 랜덤포레스트 모형으로 선정된 항공기 출발 지연 발생에 영향을 주는 기상요인들의 공항별 위험도 임계치이다. 각 변수 값에 따른 지연 확률을 모형을 통해 신출하고 확률에 따라 해당 위험도 등급과 그 때의 최솟값과 최댓값을 임계값으로 도출하였다.

15) Cut-off value : 절단값, 본 분석에서는 항공기 출발 지연 예측 모형으로부터 산출된 지연 확률의 최소값으로 정의

[표 4.5] 항공기 출발 지연에 영향을 주는 기상변수들의 위험도 임계치

영향요인	공항	위험 등급	Min	Max	영향요인	공항	위험 등급	Min	Max
10분 전 노점온도 (℃)	인천	매우 낮음	-26.8	24.6	10분 후 활주로 가시거리 (m)	인천	매우 높음	50	175
		낮음	24.7	25.8			높음	200	650
	김포	매우 낮음	-32.5	26.1			보통	700	1,400
	김해	매우 낮음	-26.8	62.4			낮음	1,500	2,100
	제주	매우 낮음	-16.1	29.5			매우 낮음	2,200	3,000
10분 전 평균풍속 (Kt)	인천	매우 낮음	1	21.0		김포	낮음	800	800
		낮음	21.1	25.5			매우 낮음	900	2,000
		보통	25.6	30.4		김해	매우 낮음	0	10,000
		높음	30.5	36.5		제주	보통	500	500
		매우 높음	36.5	38.2			낮음	600	750
	김포	매우 낮음	0	25.8			매우 낮음	800	3,000
	김해	매우 낮음	0	21.7	전운량	인천	매우 낮음	0	8
	제주	매우 낮음	0	36.5		김포	매우 낮음	0	8
10분 전 시정거리 (m)	인천	높음	0	560		김해	매우 낮음	0	8
		보통	562.5	1,650		제주	매우 낮음	0	8
		낮음	1,660	2,440	3층 운고	인천	높음	20	22
		매우 낮음	2,450	3,000			보통	25	35
	김포	낮음	0	240			낮음	50	80
		매우 낮음	245	3,000			매우 낮음	90	200
	김해	매우 낮음	0	10,000		김포	매우 낮음	20	200
						김해	매우 낮음	3	280
	제주	매우 낮음	25	10,000		제주	매우 낮음	20	230

좌측 그래프는 인천공항의 10분 전 평균 풍속에 따른 항공기 출발 지연 발생 확률(곡선)과 Cut-off value(직선) 그래프이다. 이 그래프와 같이 각 지연 영향 요인에 대하여 공항별 지연 확률에 따라 임계치와 위험도 등급을 측정하였다.



❍ 공항별 일기코드(국제기상전보식코드)의 항공기 출발 지연의 위험 등급

[그림 4.7]은 일기코드별 항공기 출발 지연의 위험 등급을 산출한 결과이다. 제주공항은 소낙눈이 보통 또는 강하게 내릴 때 일기코드 86이고 이 때, 항공기 출발 지연 확률이 약 0.74로 위험 등급이 높음을 의미한다. 제주공항에서 일기코드 86, 42, 44 이 외의 일기현상은 위험도가 매우 낮아 기상 현상으로 인한 항공기 출발 지연 확률이 Cut-off value(0.355) 이하라고 해석할 수 있다.

공항	일기코드	확률	위험등급	일기현상
제주	86	0.73699	높음	소낙눈 보통 또는 강함
	42	0.54987	보통	하늘 보이는 안개 또는 빙무 옅어짐
	44	0.42097	낮음	하늘 보이는 안개 또는 빙무 불변
김포	47	0.46653	낮음	하늘 안보이는 안개 또는 빙무 시작 또는 짙어짐
인천	75	0.97688	매우 높음	강한 눈 계속
	73	0.96502		보통 눈 계속
	85	0.93385		소낙눈 약함
	92	0.93167		관측 전 1시간 이내 뇌전 있고 관측 시에는 비 보통 또 는 강함
	71	0.92336		약한 눈 계속
	72	0.92136		보통 눈 단속
	29	0.91634		관측 전 1시간 이내 뇌전(강수유무와 무관) (10 km이상 , 현재일기현상 없음)
	97	0.89313		관측 시 강한 뇌전과 비 , 눈 , 진눈깨비 , 얼음싸라기
	45	0.87819		하늘 안보이는 안개 또는 빙무 불변
	93	0.86564		관측 전 1시간 이내 뇌전 있고 관측 시에는 눈 , 진눈깨 비, 우박 약함
	70	0.85171		약한 눈 단속
	44	0.7778	높음	하늘 보이는 안개 또는 빙무 불변
	62	0.77461		보통 비 단속
	42	0.74295		하늘 보이는 안개 또는 빙무 옅어짐
	68	0.70975		진눈깨비 약함
	22	0.70055		관측 전 1시간 이내 눈(10 km이상 , 현재일기현상 없음)
	95	0.6447	보통	관측 시 약 또는 보통 뇌전과 비 , 눈 , 진눈깨비 , 얼음싸라기
	47	0.52061		하늘 안보이는 안개 또는 빙무 시작 또는 짙어짐
	55	0.48857	낮음	강한 이슬비 계속
	50	0.48104		약한 이슬비 단속
	25	0.46702		관측 전 1시간 이내 소나기 (10 km이상 , 현재일기현상 없음)
	91	0.45885		관측 전 1시간 이내 뇌전 있고 관측 시에는 비 약함
	43	0.44267		하늘 안보이는 안개 또는 빙무 옅어짐
	46	0.38254		하늘 보이는 안개 또는 빙무 시작 또는 짙어짐
김해	44	0.84693	매우높음	하늘 보이는 안개 또는 빙무 불변

[그림 4.7] 일기코드별 항공기 출발 지연의 위험 등급

❍ 공항별 항공기 출발 지연 확률

공항별 항공기 출발 지연 확률을 살펴보면, 인천공항에서의 확률이 타 공항에 비해 상대적으로 높게 나타나고 있으나 모든 공항에서의 출발 지연 위험 등급은 매우 낮은 수준으로 공항의 영향은 미미하다고 할 수 있다.

[표 4.6] 공항별 출발 지연 확률 및 위험 등급

영향 요인	공항	확률	위험 등급
공항	제주	0.040189	< 0.355 이므로 “매우 낮음”
	김포	0.012372	
	인천	0.212344	
	김해	0.033572	

2.3.6. 결론

인천공항을 제외한 제주, 김포, 김해 공항에서 위험도와 임계치 결과가 대부분 '매우 낮음'이 나온 원인을 몇 가지 추측할 수 있다. 첫 번째는 분석에 활용된 데이터 중 정상 운항이 약 92%였기에 정상 운항과 비정상 운항(출발 지연) 데이터의 비율이 맞지 않아 모형 학습 시 정상 운항에 대하여 과적합되어 있을 수 있다. 이 경우, 정상 운항과 비정상 운항 데이터의 비율을 조정하는 샘플링 후 분석하면 모형이 개선되어 기상요인의 위험 등급 임계치의 결과가 다르게 나올 것이라고 추측된다.

두 번째, 분석을 위해 비정상 운항으로 정의한 출발 지연 데이터는 기상 원인과 기타 원인으로 인한 출발 지연이었다. 본 분석에서 기타 원인 지연 데이터가 기상 원인 데이터보다 10배 더 많이 포함되어 있었기에 지연을 발생 시키는 기상요인을 제대로 찾지 못했을 수 있다.

세 번째, 항공기 운항에 위험을 주는 위험 기상 현상이 발생한 경우 출발 지연이 아니라 결항을 하여 분석 시 반영되지 못했을 수 있다.

네 번째, 항공기 운항에 위험을 주는 기상 현상 발생 시 한 시점에 한 가지 현상만 발생하지 않았을 것이다. 여러 가지 기상 현상이 복합적으로 일어나 출발 지연을 발생시켜도 본 분석에서는 임계치 도출할 때 특정 변수를 제외한 다른 요인들은 위험도 '낮음' 이하의 상황을 고려하였다. 이러한 이유로 대부분의 위험 등급이 '낮음' 또는 '매우 낮음'이었을 것이다.

항공기 출발 지연뿐만 아니라 결항도 분석에 포함하고 복합적인 기상 현상을 고려할 수 있는 모형을 구축한다면 더 나은 성능과 정확한 임계치를 산출할 수 있을 것으로 보인다.

< 부 록 >

- 부록 1. 증발량 산출식
- 부록 2. 운량 산출식 개발
- 부록 3. 항공기 출발 지연 분석

부록 1. 증발량 산출식 개발

1-1. 기후자료 품질검사 알고리즘 기준

ASOS						AWS		
요소	물리한계검사		요소	물리한계검사		요소	물리한계검사	
	상한	하한		상한	하한		상한	하한
기온(℃)	60	-80	해면기압(hPa)	1080	500	기온(℃)	45	-35
최고기온(℃)	60	-80	최고해면기압(hPa)	1080	500	풍향(°)	360	0
최저기온(℃)	60	-80	최저해면기압(hPa)	1080	500	풍속(㎧)	75	0
이슬점온도(℃)	60	-80	습도(%)	100	0	강수유무	10	0
일강수량(㎜)	1000	0	최소습도(%)	100	0	일강수량(㎜)	1500	0
강수량(㎜)	300	0	지면온도(℃)	80	-80	습도(%)	100	0
풍향(deg)	360	0	초상온도(℃)	50	-50	기압(hPa)	1080	500
최대풍속풍향(deg)	360	0	5cm 지중온도(℃)	50	-50			
최대순간풍속풍향(deg)	100	0	10cm 지중온도(℃)	50	-50			
풍속(㎧)	100	0	20cm 지중온도(℃)	50	-50			
최대풍속(㎧)	100	0	30cm 지중온도(℃)	50	-50			
최대순간풍속(㎧)	100	0	0.5m 지중온도(℃)	50	-50			
3시간 신적설(㎝)	200	0	1.0m 지중온도(℃)	50	-50			
일 신적설(㎝)	200	0	1.5m 지중온도(℃)	50	-50			
일 적설(㎝)	200	0	3.0m 지중온도(℃)	50	-50			
일 최심적설(㎝)	2500	0	5.0m 지중온도(℃)	50	-50			
일 최심신적설(㎝)	2500	0	가조시간(hr)	15	0			
16반위 풍향	16	0	합계 대형증발량(㎝)	15	0			
현지기압(hPa)	1080	500	합계 소형증발량(㎝)	15	0			
최고현지기압(hPa)	1080	500	전운량(1/10)	10	0			
최저현지기압(hPa)	1080	500	중하층운량(1/10)	10	0			

부록 2. 운량 산출식 개발

2-1. 목측 운량 데이터 전처리 기준

번호	전운량	중하층운량	중하층운고	N	정상 여부	처리기준	운량 DS
1	0	0	0	9	비정상	맑음 → 운고 데이터 제외	삭제
2			>0	13	비정상	맑음에 중하층 운고 존재 → 운량, 운고 데이터 제외	삭제
3			NA	112,593	정상	상층운 경우 중하층운고 없음 → 운고 데이터 제외	유지
4		>0	0	0	비정상	-	-
5			>0	0	비정상	-	-
6			NA	0	비정상	-	-
7		NA	0	0	비정상	-	-
8			>0	0	비정상	-	-
9			NA	1	정상	맑음 → 운고 데이터 제외	유지
10	>0	0	0	135	비정상	상층운 → 운고 데이터 제외	삭제
11			>0	40	비정상	중하층운량 맑음인데 중하층운고 존재 → 분석 제외	삭제
12			NA	63,958	정상	상층운인 경우 중하층운고 없음 → 운고 제외	유지
13.1		>0	0	244	정상	운량 10, 10, 운고 0 →안개, 이슬비, 황사 등 시정 1Km 이하 → 운량 분석 대상 제외	삭제
13.2				3	비정상	운량 1~10, 1~9, 운고 0 → 분석 제외	삭제
14			>0	269,391	정상	-	유지
15			NA	60	비정상	운고 데이터 NA → 운고 데이터 제외	삭제
16		NA	0	0	비정상	-	-
17			>0	0	비정상	-	-
18			NA	3	정상	운량 존재하나 운고 NA → 운고 데이터 제외	유지
19	NA	0	0	1	비정상	전운량 NA -> 운량, 운고 데이터 제외	삭제
20			>0	0	비정상	-	-
21			NA	0	비정상	-	-
22		0>	0	0	비정상	-	-
23			>0	0	비정상	-	-
24			NA	0	비정상	-	-
25		NA	0	0	비정상	-	-
26			>0	0	비정상	-	-
27			NA	28,317	비정상	미관측 데이터로 운량, 운고 데이터 제외	삭제
계				474,768	-	-	-

2-2. 천리안 위성 운량 데이터 전처리 기준

번호	전운량	운정고도	N	정상 여부	처리기준	운량 DS
1	0	0	0	비정상	-	삭제
2		>0	438	비정상(32767 값 이외)	운량, 운고 데이터 제외	삭제
		>0	43,501	정상(32767 값)	"32767" → 청천영역(맑음) → 운정고도 0으로 대체	유지
3		NA	77	정상	운정고도 0으로 대체 → 운고 데이터 삭제	유지
4	>0	0	0	비정상	-	-
5		>0	392,906	정상(32767 값 이외)	-	유지
		>0	36,439	비정상(32767 값)	"32767" → 청천영역(맑음) → 운정고도 0으로 대체	삭제
6		NA	791	비정상	운량, 운고 데이터 제외	삭제
7	NA	0	0	비정상	-	-
8		>0	35	비정상(32767 값 이외)	운량, 운고 데이터 제외	삭제
		>0	49	비정상(32767 값)	운량, 운고 데이터 제외	삭제
9		NA	532	비정상	운량, 운고 데이터 제외	삭제
계			474,768	-	-	-

2-3. 운고운량계 운량 데이터 전처리 기준

번호	운량	하층운고	N	정상 여부	처리기준	운량 DS
1	0	0	0	비정상	-	삭제
2		>0	4,894	비정상(7620값 이외)	운량, 운고 데이터 제외	삭제
		>0	174,084	정상(7620 값)	“7620" → 구름없음(맑음) → 하층운고 “0”으로 대체	유지
3		NA	0	정상	-	-
4	>0	0	0	비정상	-	-
5		>0	198,380	정상(7620 값 이외)	-	유지
		>0	59,224	비정상(7620 값)	“7620" → 구름없음(맑음) → 하층운고 “0”으로 대체	삭제
6		NA	0	비정상	-	-
7	NA	0	1	비정상	운량, 운고 데이터 제외	삭제
8		>0	79	비정상(7620 값 이외)	운량, 운고 데이터 제외	삭제
		>0	838	비정상(7620 값)	“7620" → 구름없음(맑음) → 하층운고 “0”으로 대체	삭제
9		NA	37,268	비정상	운량, 운고 데이터 제외	삭제
계			474,768	-	-	-

2-4. 히마와리 위성 및 기상 관측 정보 분석 시 운량 데이터 전처리 기준

번호	전운량	중하층운량	운형	N	N %	정상 여부	처리기준	운량 DS	비고
1	-9	-9	구름없음(-)	10,335	16%	비정상	목측운량 없음 → 운량 데이터 삭제	삭제	-
2			구름있음	1	0%	비정상			
3		0	구름없음(-)	1,951	3%	비정상			
4			구름있음	0	0%	비정상			
5		>0	구름없음(-)	1,585	2%	비정상			
6			구름있음	0	0%	-			
7	0	-9	구름없음(-)	0	0%	-	-	-	-
8			구름있음	0	0%	-	-	-	-
9		0	구름없음(-)	8,170	13%	정상	-	유지	-
10			구름있음	1	0%	비정상	운량 0, 구름있음 → 운량 데이터 삭제	삭제	-
11		>0	구름없음(-)	0	0%	정상	-	-	-
12			구름있음	0	0%	-	-	-	-
13	>0	-9	구름없음(-)	0	0%	-	-	-	-
14			구름있음	2	0%	정상	-	유지	-
15		0	구름없음(-)	0	0%	-	-	-	-
16			구름있음	7,568	12%	정상	-	유지	-
17		>0	구름없음(-)	972	2%	비정상	안개에 의한 운량 >0 → 운량데이터 삭제	삭제	
18			구름있음	33,453	52%	정상	-	유지	-
계				64,038	100%	-	-	-	-

※ 목측 전운량의 결측(-99)인 4건을 제외한 수치임

부록 3. 항공기 출발 지연 분석

3-1. 항공기상 통계 품질검사 기준 (항공기후통계 지침)

AMOS 기상 데이터

풍향	▪ 0 ~ 360
풍속	▪ 0 ~ 50 kt
시정거리	▪ 0~10,000 m
활주로 가시거리	▪ 0~10,000 m
기압	▪ 960 ~ 1045 hPa
습도	▪ 0 ~ 100
기온	▪ 1,2,12월 : -25 ~25 ℃ ▪ 3,4,10,11월 : -15 ~35 ℃ ▪ 5~9월 : 0 ~40 ℃
강수량	▪ 1,2,12월 : - ▪ 3,4,10,11월 : 0 ~ 70 mm ▪ 5~9월 : 0 ~140 mm

구름

▪ 운량 : 0~10
▪ 운고 : 0~40,000ft
▪ 운형 0 ~9

▪ 전운량 = 0 : 1,2,3,4, 층 운량, 운형, 운고 공란
▪ 전운량 〉 0

운고	- 1층 〈 2층 〈 3층 〈 4층
운량	- 1층 〈 2층 ≤ 3층 ≤ 4층 - 전운량 ≥ 최상층 운량 - 1층운만 관측될 경우, 최상층운량=전운량 - 2층운 이상 관측될 경우, - 8 ≥ 1층+2층+3층+4층 ≥ 전운량 ≥ 최상층 운량
운형	- 운형 6,8,9 일때, 운고 ≤ 6500 ft - 운형 7 일때, 운고 ≤ 2000 ft - 운형 3,4 일때, 6500 ft ≤ 운고 ≤16000 ft - 운형 5 일때, 5000 ft ≤ 운고 ≤ 10000 ft - 운형 0,1,2 일때, 16000 ft ≤ 운고 ≤ 40000 ft

참고문헌

국가기상위성센터(2016). “히마와리 8 위성 관측자료를 이용한 RGB 컬러합성기법 매뉴얼”, 국가기상위성센터 위성분석과 기술노트 2016-1호

국가기상위성센터(2014). “천리안 위성과 지상 목측 운량 비교 결과”

국가기상위성센터(2012). “CP 알고리즘 기술 분석서(Algorithm Theoretical Basis Document)”

국가기상위성센터(2012). “CA 알고리즘 기술 분석서(Algorithm Theoretical Basis Document)”

기상기술융합팀(2016). “청내 빅데이터 분석 과제 수요 조사 계획”

관측정책과(2016). “빅데이터 분석과제 신청서”

관측정책과(2016). “증발량 관측방법 개선을 통한 자동화 추진 계획”

나득균 외(2014). “고창표준기상관측소 증발량 비교관측 결과”, 관측기반국 관측정책과 기술노트 2014-2호

유상진 외(2015). “천리안 위성자료를 이용한 운량산출 기술 개선”, 국가기상위성센터 위성분석과 기술노트 2015-4호

이병일 외(2007). “MTSAT-1R 정지기상위성 자료를 이용한 전운량 산출 알고리즘 개발”, 대기 vol.17

이영종 외(2014). “시뮬레이션 기반의 지상지연 프로그램 적용방안에 관한 연구 - 제주국제공항을 사례로”, 한국항공운항학회지 vol.23

임창수, 송주일(2010). “지역특성에 따른 기후변화가 증발에 미치는 영향분석”, 국제수문개발계획(IHP) 제7단계 제2차년도

임창수(2011). “증발량 산정을 위한 입사태양복사식 비교”, 농업과학연구 vol.38

항공기상청(2015). “항공기상업무지침(제9차 개정판)”

항공기상청(2016). “항공기후통계지침”

허국강(2010). “항공기 비정상 운항의 효율적 관리방안에 관한 연구”, 항공진흥 2010년 제2호

찾아보기

[증발량 산출식 개발]

결측치, 7
모형설명력(R^2), 15
이상치, 7
일반선형모형(GLM), 13
증발량, 5
Box plot, 8
K-fold cross validation, 16
kNN Imputation, 8
PM 산출식, 6
Random Forest, 13
SMAPE, 16

[운량 산출식 개발]

격자점수, 29
결측치, 29
교호작용, 25
다항 로지스틱 회귀모형, 26
로직체크, 23
순서형 로지스틱 회귀모형, 26
이상치, 29
장파복사, 21
회귀계수, 26
accuracy, 32
depth, 32
ntree, 32
Random Forest, 31
Stepwise 변수 선택법, 25

[항공기 출발 지연 분석]

결측치, 39
다중공선성, 41
이상치, 39
임계치, 28
정확도, 43
AUC, 43
Cut-off value, 44
Gini, 43
kNN Imputation, 40
LogLoss, 43
Random Forest, 43
QFE, 39
QNH, 39

2016년 정책 목표

영향예보로의 전환을 통한 기상재해 리스크 경감

기상청 기상서비스진흥국 기상융합서비스과