응용통계연구 (2011)
24(3), 535–546
DOI: 10.5351/KJAS.2011.24.3.535

남한지역 겨울철 황사출현일수에 대한 범주 예측모형 개발

손건태[1] · 이효진[2] · 김승범[3]

[1]부산대학교 통계학과, [2]부산대학교 통계학과, [3]국립기상연구소 황사연구과

(2011년 3월 접수, 2011년 4월 채택)

요 약

본 연구는 겨울철 남한지역 황사출현일수에 대한 이 범주 계절예측모형 개발을 목적으로 수행되었다. 최근 31년간 관측된 황사출현일수를 예측량으로 하고, 황사발원지 기상요소(지상기온, 강수량, 강설량, 지상풍속)에 대한 NCEP 재분석자료 예측치와 광역규모 기후지수들을 잠재적 예측인자로 사용하였다. 월별로 구분하여 예측모형을 개발하기 위하여 네 종류 통계모형(중회귀모형, 로지스틱 회귀모형, 의사결정나무모형, 지지벡터기계)을 각각 적용하였다. 예측모형 평가측도인 정분류율, 탐지확률, 잘못된 경고를 사용하여 모형 비교하고 예측모형을 제안하였다.

주요용어: 황사출현일수, 범주예보, NCEP재분석자료, 기후지수.

1. 서론

황사(Asian dust)현상은 중국북부와 몽골의 건조지역에서 발생하고(dust emission), 장·단거리를 이동하여(dust transport), 서서히 낙하하는 현상(dust deposition)을 말한다. 황사발생은 넓은 영역에 분포된 황사발원지에서 저기압에 동반된 한랭전선 후면의 강풍대가 접근하게 되면서 지표면으로부터 날려진 먼지입자가 강한 바람과 함께 상공으로 비산하여 이루어진다. 황사발원지에서 발생한 황사량의 30% 정도는 발원지에 재침적되고, 20% 정도는 주변지역으로 수송되며, 나머지는 장거리 수송되어 북동아시아뿐만 아니라 태평양 건너 북미대륙까지 영향을 미친다. 황사의 성분은 규소와 니켈 등 광물성 물질과 유기물과 황산염 등으로 이루어지며, 국가 간의 오염 확산이라는 측면에서 매우 중요하게 취급되고 있다.

동북아시아 지역은 경제발전 속도가 빠른 지역으로서, 한반도로 날아오는 황사는 중국의 북동부 공업지대에서 방출되는 중금속 오염물질까지 함께 섞여 수송되므로 더욱 심각하게 인식되고 있다. 황사는 시정 불량, 호흡기질환 급증, 휴교, 태양에너지의 감쇄, 정밀기계 및 반도체의 손상 등 첨단산업의 피해 등 경제,사회적으로 큰 피해를 주므로 기상재해로서 인식되었다. 이에 따라 황사현상 예측의 중요성이 점차 증대되고 있다. 기상청은 2002년부터 황사특보제를 시행하고 있으며, 2008년부터는 황사출현일수(Asian dust days)에 대한 객관적 계절예보를 위하여 통계적 예측모형을 개발하여 봄철과 겨울철 기후전망에 활용하고 있다.

황사예측에 대하여 주로 단기예측에 중점을 두고 연구가 이루어져 왔다. 윤순창과 박경선 (1991)은 엔트로피 궤적에 의한 황사의 장거리 이동경로 분석을, 정용승과 김태균 (1991)은 장거리 이동에 대한 사례연구를 발원지 추적과 함께 연구하였다. 전종갑 등 (2000)는 한반도에서 관측된 황사의

본 연구는 국립기상연구소 2010년도 용역연구과제 “독자 전지구 황사예측시스템 구축 및 황사계절예측모형 개발의 지원”으로 수행되었습니다.
[1]교신저자: (609-735) 부산시 금정구 장전동, 부산대학교 통계학과, 교수. E-mail: ktsohn@pusan.ac.kr

특성 및 장거리 수송패턴을 분석하였다. 기상청은 황사농도에 대한 단기예측을 위하여 수치모형인 ADAM(Asian Dust Aerosol Model)을 개발하여 2002년부터 현업예보에 활용하고 있다. 이를 위하여 박순웅 (2002)은 봄철 WMO 종관 관측소자료를 이용하여 황사발생지역 분포, 발생빈도, 발생 토양 종류별 임계마찰속도를 규명하고 에어로졸 수송모형을 개발하였으며, Park과 In (2003)은 2002년 3월 황사에 대한 ADAM 결과를 분석하였다. ADAM은 지역모형을 바탕으로 이루어져 있으며 대기모형과 결합된 모형이 아닌 off-line 수송모형이기 때문에 황사의 장기예측에는 활용하기 어려운 형편이다.

전 세계적으로 황사예측을 위하여 운용되는 수치모형은 단기예측을 대상으로 하고 있으나, 황사장기예측을 위한 수치모형의 개발은 현재 초기단계에 있다. 따라서 본 연구에서는 관측치 또는 재분석자료를 사용하는 통계모형을 적용하기로 하였다. 계절예측을 위하여 황사발원지에서 황사발생의 원인이 되는 예측인자를 고려하였으며, 황사이동에 영향을 주는 광역 기후패턴을 고려하기로 하였다. 황사발원지에 대한 연구로는 Lim과 Chun (2006)은 북동아시아지역의 황사특성에 대하여 황사발원지를 3지역(건조지역, 반건조지역, 경작지역)으로 구분하여 분석하였다. Zhang 등 (2008)는 중국 북서지역의 황사발원지를 5개 지역(내몽골 중앙지역, 몽골 남서부의 고비지역, 몽골 남부와 내몽골 북부의 고비지역, 황하 중류의 사막지역, 타클라마칸 사막지역)으로 구분하여 이들이 중국과 몽골의 황사발생에 미치는 영향을 조사하였다. Tian 등 (2007)은 일본 황사출현빈도와 중국북부 먼지폭풍 빈도와의 관계를 관측치와 NCEP 재분석자료를 사용하여 조사하였다.

남한지역에서 관측된 황사출현일수를 볼 때, 봄철에 90%가까이 발생하고 있어, 봄철에 대한 황사예측에 중점을 두고 있다. 그러나 연간 겨울철 황사출현일수의 변화를 보면 1997년까지 겨울철 황사출현이 거의 없다가 이후 급격히 증가하고 있으며, 2010년 11월에는 관측 이래 첫 황사특보가 발령되는 등 겨울철 황사현상에 대한 관심이 커지고 있다.

본연구는 겨울철 남한지역 황사출현일수에 대한 이 범주 계절예측모형 개발을 목적으로 하였다. 이 범주 분류를 위하여 최근 30년간 관측된 황사출현일수의 평균과 표준편차를 사용하였으며, 각 년도의 황사출현일수가 (평균 + 0.5 × 표준편차)보다 적으면 '평년수준(normal)'으로 분류하고, 같거나 많으면 '평년수준보다 많음(above normal)'으로 분류하였다. 본 연구에서 황사발원지는 Lim과 Chun (2006)에서 구분된 세 지역(건조지역, 반건조지역, 경작지역)을 고려하였으며, 황사발원지 기상요소로 지상기온, 강수량, 강설량, 지상풍속을 고려하였다. 황사발원지는 기상관측소가 거의 없는 매우 넓은 영역으로 이루어져 있어, 관측치를 얻기 어려우므로 미국 국립대기과학연구소 환경예측센터(NCEP/NCAR)에서 제공하는 재분석자료(reanalysis data)를 사용하였다.

김연희 등 (2008)는 기후시스템을 구성하는 각 권역내의 변동과 권역간 다양한 상호작용은 다양한 시공간 규모의 기후변동을 유발하며, 이러한 기후변동은 광역규모 기후지수(large-scale climate index)에 의하여 효율적으로 나타낼 수 있다고 하였다. 기후지수들은 대규모 순환패턴에 따라 이루어지는 기후변동의 이해와 다양하고 넓은 지역의 기후감시를 목적으로 전 세계적으로 연구되고 있다. 남한지역의 황사현상은 장거리 이동에 따른 결과이므로 광역규모 기후지수들을 고려하였다. 기후지수들은 미국 국립해양 대기청(National Oceanic and Atmospheric Administration; NOAA)의 기후 진단 센터(Climate Diagnostic Center; CDC)와 기후 예측 센터(Climate Prediction Center; CPC)에서 제공하는 기후지수들 중 16 종류의 기후지수를 연구에 사용하였다.

상관분석 결과 월별(겨울철, 11월, 12월, 1월, 2월)로 구분하여 각각 예측모형을 개발하였다. 11월은 가을에 속하지만 6월부터 10월은 황사현상이 없으므로 겨울철에 포함시켜 분석하였다. 이 범주 예측치 생산을 위하여 세 가지 방법을 적용하고 결과를 비교하여 최적 예측모형을 제안하고자 하였다. [방법 1]은 계량치 예측모형(중회귀모형)을 적합하고 생산된 계량 예측치에 분류기준을 적용하여 이 범주 예측치를 생산하고, 확률 예측모형(로지스틱 회귀모형, 의사결정나무모형)을 적합하여 생산된 확률 예

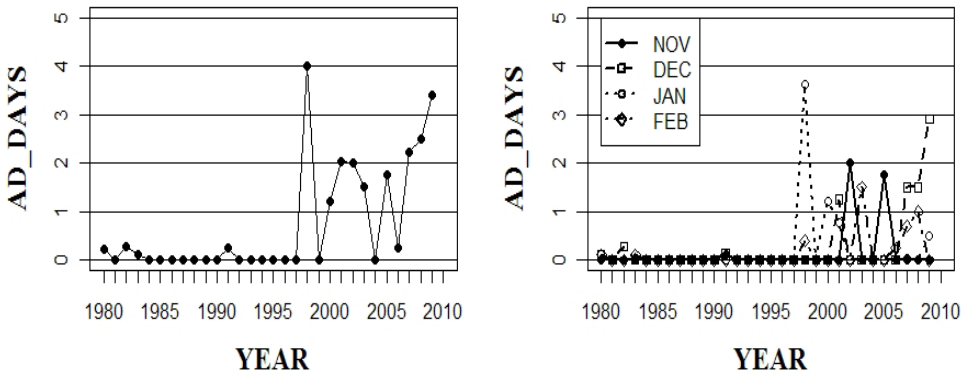

그림 **2.1.** Time series plot of Asian dust days in the Winter season

측치에 문턱치(threshold)를 고려하여 이 범주 예측치를 생산하고, 범주 예측모형(지지벡터기계)을 적용하여 이 범주 예측치를 생산하였다. 각 모형의 이 범주 예측결과는 2×2 분할표로 정리하였으며, 최적 예측모형을 결정하기 위하여 예측모형 평가측도인 정분류율(hit rate), 탐지확률(probability of detection), 허위경고율(false alarm rate)를 비교하였다. 모형검증을 위하여 교차검증(leave-one-out cross validation)을 적용하였다.

Sohn 등 (2008)은 남한지역 봄철 황사출현일수에 대한 계량치예측, 확률예측, 세 범주예측을 위하여 황사발원지 기상요소에 대한 NCEP 재분석자료를 예측인자로 사용하였다. Sohn 등 (2009)은 호남지역 대설특보를 위하여 로지스틱 회귀모형과 신경회로망을 적용하였다. Sohn과 Park (2008)은 이 범주 예보를 위한 문턱치 결정을 위하여 예측성 평가측도를 활용하는 가이던스를 제안하였다. 로지스틱 회귀모형은 순서형 자료를 예측량으로 하는 확률예측모형으로 유용하게 사용된다. 의사결정나무는 Breiman 등 (1984)에 의하여 CART(Classification And Regression Tree)로 제안되어 확률예측과 계량치 예측을 위하여 사용되는 모형이며, 지지벡터기계(Support Vector Machine; SVM)는 Vapnik (1996)에 의하여 제안된 이 범주 분류모형으로 다 범주로 확장이 가능하며, 많은 분야에서 적용되는 모형이다. 자료분석과 모형개발을 위하여 SAS/E-Miner와 R 패키지의 library(e1071)의 svm함수를 사용하였다.

2. 자료와 방법론

2.1. 황사출현일수

황사출현일수에 대한 예보모형 개발을 위하여 남한지역 28개 지점의 월별 황사출현일수의 공간적 평균치(앞으로 단순히 황사출현일수로 명명함)를 목적변수(target variable, predictand)로 사용하였다. 자료기간은 30년(1980년~2009년)이다. 겨울철과 월별(11월, 12월, 1월, 2월)로 구분하여 각각 황사출현일수 예측모형을 개발하고자 하였다. 그림 2.1은 겨울철 황사출현일수의 시계열그림이다. 겨울철은 1998년까지 거의 발생하지 않다가 1999년부터 급격히 증가하고 있음을 알 수 있다. 겨울철의 변화를 보면 1980년대에는 0.06일, 1990년대는 0.425일, 2000년대에는 1.69일로 점차 증가하고 있으며, 표준편차의 변화를 보면 1980년대에는 0.107일, 1990년대는 1.259일, 2000년대에는 1.016일로 90년대 이후에는 80년대와 달리 변동성도 커졌다. 또한 각 월별로 구분된 황사출현일수의 시계열을 보면 매우 다르게 변화하고 있다.

30년간 황사출현일수에 대한 기초 통계치는 표 2.1과 같다. 겨울철의 황사출현일수 평균와 표준편차는

표 **2.1.** Basic statistics of Asian dust days

Month	N	Mean	STD	Frequency	
				Normal	Above normal
Winter	30	0.725	1.148	22	8
NOV	30	0.130	0.476	28	2
DEC	30	0.256	0.660	26	4
JAN	30	0.197	0.687	27	3
FEB	30	0.131	0.339	26	4

표 **2.2.** Correlation analysis among months

	NOV	DEC	JAN
DEC	−0.102	1	0.006
	(0.592)		(0.974)
JAN	−0.088	0.006	1
	(0.645)	(0.974)	
FEB	−0.109	0.499	0.088
	(0.566)	(0.005)	(0.636)

각각 0.725일과 1.148일이다. 각 월의 황사출현일수에 대한 이 범주(normal, above normal)는 30년간 황사출현일수 평균과 표준편차를 사용하여 다음과 같은 기준으로 분류하였으며, 월별로 각 범주에 속한 도수를 표 2.1에 정리하였다.

분류기준 : (평균 + 0.5 × 표준편차)보다 적으면 normal, 같거나 많으면 above normal.

월별 황사출현일수에 대한 상관분석 결과를 표 2.2에 요약하였다. (12월, 2월)을 제외하면 상관성이 유의하지 않으므로 월별로 예측모형을 개발하기로 하였다.

2.2. 황사발원지 기상요소에 대한 NCEP/NCAR 재분석자료

황사 발원지는 매우 넓은 지역에 분포되어 있고, 황사발생은 식생과 토양의 특성 및 기상상태에 영향을 받으므로 몇 개의 영역으로 구분할 필요가 있다. 본 연구에서는 그림 2.2와 같이 Lim과 Chun (2006)에서 제시된 세 영역을 고려하였다. A 영역은 (100°E∼110°E, 35°N∼45°N)으로 건조지역인 고비지역을 포함하고 있고, B 영역은 (110°E∼120°E, 40°N∼45°N)으로 반건조지역인 내몽골지역을 포함하고 있으며, C 영역은 (120°E∼125°E, 40°N∼50°N)으로 경작지역인 만주지역을 포함하고 있다.

30년간(1979년∼2008년)의 세 영역에 대한 지상기온(℃), 강우량(mm), 강설량(mm), 지상풍속(m/s)에 대한 NCEP/NCAR 재분석자료의 영역별 월평균 자료를 수집하였다. 본 연구에서는 황사발생원인에 대한 해석을 위하여 재분석자료의 예측치를 예측인자로 사용하기로 하였다. 즉, 다음해 겨울철에 황사발원지의 기상상태를 설명하기 위하여 다음해에 해당하는 재분석자료의 예측치를 예측인자로 사용하기 위함이다. NCEP/NCAR 재분석자료의 예측치 생성을 위하여 황사발원지별 기상요소별 12종류(3영역 × 4요소)에 대하여 각각 계절형 자기회귀-누적-이동평균모형을 적합하였으며, 다음의 결과를 얻었다.

첫째, 광학분석(spectral analysis)을 수행한 결과 모든 기상요소에서 주기 12개월이 강하게 검출되었다. 둘째, 자기상관함수, 편자기상관함수, 식별통계량(ESACF)을 사용하여 모형을 식별하였으며, 추정된 모형식은 표 2.3에 요약하였다. 표 2.3에서 B는 후향연산자이며, a_t는 오차항이고, AT는 A영역 지상기온, BT는 B영역 지상기온을 의미한다.

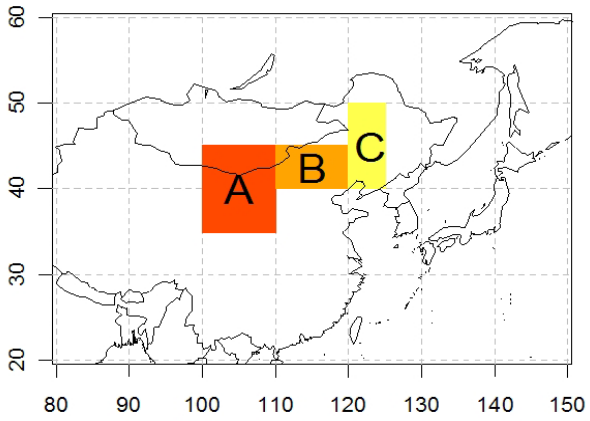

그림 **2.2.** Source regions of Asian dust

표 **2.3.** Estimated models for each factor in source regions, (T: Temperature, P: Precipitation, S: Snowfall, WS: Wind Speed)

Region	Factor	Estimated model
A	T	$(1-0.19557B)(1-B^{12})AT_t = (1-0.80152B^{12})a_t$
	P	$(1-B^{12})AP_t = (1-0.69423B^{12})a_t$
	S	$(1-0.32068B)(1+0.58498B^{12})(1-B^{12})AS_t = (1-0.49828B^{24})a_t$
	WS	$(1-B^{12})AWS_t = (1-0.72344B^{12})a_t$
B	T	$(1-0.2115B)(1-B^{12})BT_t = (1-0.81654B^{12})a_t$
	P	$(1-B^{12})BP_t = (1-0.71098B^{12})a_t$
	S	$(1-0.50079B)(1-B^{12})BS_t = (0.62116B^{12})a_t$
	WS	$(1-0.10293B^2-0.13377B^3)(1-B^{12})BWS_t = (1-0.78748B^{12})a_t$
C	T	$(1-0.25816B)(1-B^{12})CT_t = (1-0.85393B^{12})a_t$
	P	$(1-0.1966B)(1-B^{12})CP_t = (1-0.76263B^{12})a_t$
	S	$(1-0.5661B)(1-B^{12})CS_t = (1-0.59941B^{12})a_t$
	WS	$(1-B^{12})CWS_t = (1-0.80079B^{12})a_t$

셋째, 예측오차에 대한 자기상관성 조사를 수행한 결과 모든 경우에서 자기상관성이 유의하지 않은 것으로 나타났다.

넷째, 종합적으로 볼 때 12 종류 예측 모형식들은 모두 비교적 높은 예측성을 지니고 있다고 판단되어, 예측치들을 황사출현일수 계절예측모형의 잠재적 예측인자로 사용하기로 하였다.

2.3. 광역규모 기후지수

남방진동지수(Southern Oscillation Index)와 북극진동(Arctic Oscillation)과 같은 광역규모 기후지수들은 대규모 순환패턴에 따라 이루어지는 기후변동의 이해와 다양하고 넓은 지역의 기후감시를 목적으로 전 세계적으로 연구되고 있으며, 일반적으로 기후지수와 지역기후와의 상관분석을 활용하고 있다. 본 연구에서는 황사운송과 관련성이 있는 기후지수를 찾고자 하였다. NOAA의 CDC와 CPC에서 제공하는 16종류 기후지수들 표 2.4를 연구에 사용하였다. 기후지수들은 월 자료로 구성되어 있다. 광역규모 기후지수들에 대하여 시계열분석을 수행한 결과 유의한 주기가 다수 있는 경우도 있고, 예측 결과가 좋은 지수들도 있으며, 좋지 않은 경우도 있어 광역규모 기후지수는 관측치를 그대로 사용하기로 했다.

표 **2.4.** List of the climate indices updated on a monthly basis from the Climate Diagnostic Center and Climate Prediction Center, NOAA

Abbr.	Full name	Source
AMO	Atlantic Multidecadal Oscillation	CDC
AO	Arcric Oscillation	CPC
ESO	Equatorial SOI	CPC
GML	Global Mean Land Ocean Temperature Index	CDC
MEI	Multivariate ENSO Index	CDC
NAO	North Atlantic Oscillation	CPC
NOI	Northern Oscillation Index	CDC
ONI	Oceanic Nino Index	CPC
PDO	Pacific Decadal Oscillation Index	CDC
PNA	Pacific / North American Pattern	CPC
SOI	Southern Oscillation Index	CPC
SWM	SW Monsoon Region Rainfall	CDC
TNA	Tropical Northern Atlantic Index	CDC
TSA	Tropical Southern Atlantic Index	CDC
WHW	Western Hemisphere Warm Pool	CDC
WPO	West Pacific Oscillation	CPC

2.4. 이차 잠재적 예측인자

황사출현일수에 대한 예측모형개발을 위하여 잠재적 예측인자는 NCEP/NCAR 재분석자료의 예측치(3영역 × 4요소 × 12개월 = 144종류)와 광역규모 기후지수(16종류 × 12개월 = 192종류)로 이루어진 총 336종류 예측인자로 구성된다. 모든 변수를 함께 사용하여 모형을 개발하기에는 어려우므로 황사출현일수와 각각의 영역별 기상요소별 예측인자들 사이의 상관분석을 수행하여, 유의수준 0.05에서 유의한 상관이 있는 변수를 선택하고, 상관계수 부호를 고려하여 이차 잠재적 예측인자들을 생성한다. 선택된 이차 잠재적 예측인자는 표 2.5에서 보듯이 겨울철에 대하여 재분석자료 예측치는 17개, 광역규모 기후지수는10개로 총 27개, 11월은 재분석자료 예측치가 5개, 광역규모 기후지수는 4개로 총 9개, 12월은 재분석자료 예측치10개, 광역규모 기후지수8개로 총 18개, 1월은 재분석자료 예측치와 광역규모 기후지수가 각각 7개로 총 14개, 2월은 재분석자료 예측치가 10개, 광역규모 기후지수가 5개로 총 15개다.

표에서 FAT4는 'A지역 지상기온(T)의 재분석자료 예측치(F) 중 4월 자료'를 의미한다. 즉, F는 예측치를 뜻하고, 세 지역(A, B, C)과 기상요소(T, P, S, WS)와 월로 구성되어 있다. NAO5는 '지난해의 5월 북대서양진동 지수(NAO)'를 의미한다. 월별로 서로 다른 기후지수가 선정되었다. 이는 남한지역 황사출현일수가 월별로 매우 다른 원인에 의하여 영향을 받는 것으로 해석될 수 있어, 이에 대한 향후 연구가 필요하다고 판단된다.

2.5. 통계모형과 이 범주예측치 생성 전략

황사출현일수에 대한 예측모형개발을 위하여 이차 잠재적 예측인자를 사용하여 세 가지 방법(계량치예측모형 사용, 확률예측모형 사용, 범주예측모형 사용)으로 범주예측모형을 개발하였다. 계량치 예측모형으로 중회귀모형을 적용하였다. 지연상관분석에서 생성된 이차 잠재적 예측인자를 예측인자로 적용하고, 변수선택은 앞으로부터 단계별 회귀방법을 적용하였다. '평균 + 0.5 × 표준편차' 분류기준에 따라 이 범주 예측치를 생산하였다. 확률예측모형으로 로지스틱 회귀모형과 의사결정나무를 각각 적용하였다. 이 범주 로지스틱 회귀모형과 의사결정나무는 이항확률값을 생산한다. above normal이 normal보

표 **2.5.** Secondary potential predictors

Month	Secondary potential predictors	
	NCEP reanalysis data	Climate Indices
Winter	FAT1, FAT11, FBT1, FBT11, FCT3, FAP1, FCP3, FAS1, FBS1, FBS10, FCS1, FCS10, FAWS11, FBWS10, FBWS11, FCWS10, FCWS11	AMO7, AO7, GML7, NAO7, PDO7, SWM8, TNA10, TSA12, WHW7, WPO2
NOV	FBS1, FCS2, FAWS3, FBWS11, FCWS3	AMO1, GML1, TNA1, WPO7
DEC	FAT12, FBT1, FAP1, FBP11, FCP3, FAS12, FBS12, FCS12, FBWS11, FCWS11	AO7, GML7, NAO7, NOI4, PDO4, PNA9, NAO9, SOI2
JAN	FAT3, FBT11, FCT11, FAS3, FBS3, FCS10, FAWS4	AMO6, GML7, NAO6, NOI8, SWM10, WHW2, WPO9
FEB	FCT2, FAP1, FCP3, FAS2, FBS2, FCS2, FBWS3, FBWS11, FCWS5, FCWS11	AMO10, GML10, SOI12 TNA10, WPO7

표 **2.6.** 2×2 table

Month	Secondary potential predictors		Total
	normal	above normal	
normal	D (negative correction)	B (false alarm)	B + D
above normal	C (miss)	A (hit)	A + C
Total	C + D	A + B	A + B + C + D

다 빈도수가 매우 작기 때문에 확률예측보다 문턱치를 고려한 범주예측이 선호된다. 즉, above normal일 확률이 문턱치보다 같거나 크면 above normal이라고 예측한다. 최적 문턱치는 0과 1사이에서 0.1씩 변화시키며 2×2 분할표를 생성하여 예측결과가 가장 좋다고 판단되는 문턱치로 결정하였다. 의사결정나무에서는 엔트로피 지수를 분리기준으로 사용하였다. 범주예측모형은 지지벡터기계(SVM)를 적용하였다. 커널은 방사기저함수(radial basis function)를 사용하였으며, 감마값은 각 설명변수 수의 역수로 사용하였다.

2.6. 통계모형과 이 범주예측치 생성 전략

네 모형의 이 범주 예측 결과를 비교하여 최적 모형을 범주 예측모형으로 제안하고자 하였다. 예측 결과는 표 2.6의 2×2 분할표로 요약하고, 아래 식으로 정의되는 예측성 평가측도(skill score)인 정분류율(HR), 탐지확률(POD), 허위경고율(FAR)을 구하여 비교하였다. 당연히 above normal이던지 normal이던지 모든 경우에 잘 맞춰야 좋은 예측모형이다. 황사는 above normal인 경우의 피해가 normal보다 크므로 더 중시하게 된다. 황사출현일수 경우 normal의 발생율이 90% 정도로 크기 때문에 언제나 normal이라고 예보하는 경우 HR이 90% 정도 높게 나타나지만, above normal을 예측하지 못하는 큰 문제가 있다. POD는 관측범주가 above normal인 경우 above normal이라고 예측한 비율이므로 above normal에 중점을 둔 측도이다. 그러나 언제나 above normal이라고 예보한다면 POD는 최대값인 1을 갖지만 FAR도 1이 되어 예측모형으로 사용할 수 없다. FAR은 예보관이 above normal이라고 예보한 경우에 허위경고율이므로 FAR이 작을수록 좋은 모형이 된다. 따라서 HR과 POD는 가능한 크고 동시에 FAR은 가능한 작은 결과를 보이는 예측모형을 선택했다.

표 **3.1.** Skill scores of forecast models for each month

Month	Model(Threshold)	Whole data(%)			CV
		HR	POD	FAR	
Winter	REG	93.3	87.5	12.5	93.3
	Logistic REG(0.7)	96.7	87.5	0.0	90.0
	Decision TREE(0.5)	96.7	87.5	0.0	96.7
	SVM	96.7	100.0	11.1	93.3
NOV	REG	93.3	100.0	50.0	83.3
	Logistic REG(0.1)	93.3	100.0	50.0	90.0
	Decision TREE(0.5)	93.3	0	*	93.3
	SVM	100.0	100.0	0.0	93.3
DEC	REG	73.3	75.0	70.0	70.0
	Logistic REG(0.7)	93.3	50.0	0.0	90.0
	Decision TREE(0.5)	93.3	100.0	33.3	90.0
	SVM	96.7	75.0	0.0	86.7
JAN	REG	100.0	100.0	0.0	96.7
	Logistic REG(0.6)	96.7	100.0	25.0	89.7
	Decision TREE(0.5)	93.3	100.0	40.0	90.0
	SVM	100.0	100.0	0.0	86.7
FEB	REG	86.7	100.0	50.0	83.3
	Logistic REG(0.2)	83.3	100.0	55.6	79.3
	Decision TREE(0.5)	86.7	0.0	*	86.7
	SVM	93.1	50.0	0.0	86.2

모형검증은 자료의 수가 적으므로 교차검증(leave-one-out cross validation)을 적용하였다. 교차검증의 결과도 2×2 분할표로 요약하여 정분류율, 탐지확률, 허위경고율을 구하고, 전체자료 사용 경우와 비교하여 모형의 안정성을 판단하였다.

3. 겨울철 월별 황사 계절예측모형 개발결과

선택된 이차 잠재적 예측인자를 사용하여 황사출현일수에 대한 예측모형개발 결과는 표 3.1에 예측모형에 사용된 이차잠재적 예측인자는 표 3.2에 정리하였다. 표 3.1은 30년 전체자료를 사용한 경우의 예측성 평가측도(HR, POD, FAR), 교차검증(CV) 경우의 HR, 최종 선택된 예측인자(재분석자료 예측치, 광역규모 기후지수)로 구성되어있다. 표 3.2에서 *는 해당 예측인자에 유의한 변수가 없다는 뜻이며, INT는 절편(또는 상수항)이 유의하여 모형에 사용된 경우를 뜻하고, T는 시간(년)을 뜻하며, ALL은 이차 잠재적 변수 전체를 사용한 것을 뜻한다. 겨울철 전체와 월별 최종 유의인자들이 매우 다르게 선택된 것을 알 수 있다. 이는 월별로 다른 기후패턴과 기상요소들이 영향을 주고 있기 때문이라 생각된다.

겨울철에 대하여 네 가지 모형을 적합한 결과 계산된 HR, POD, FAR, CV를 볼 때 모두 유용한 예측모형이라고 판단되며, 그 중 의사결정나무가 가장 좋은 결과를 제공하고 있다. SVM은 POD가 100%이나 교차검증 결과는 83.8%로 모든 데이터를 사용한 경우보다 상대적으로 낮게 나타났다. 따라서 겨울철 황사출현일수에 대한 이 범주 예보는 다음의 식으로 구성되는 의사결정나무를 예측모형으로 제안한다. 겨울철의 의사결정나무에는 재분석자료 예측치는 사용되지 않으며, 광역기후지수인 GML7이 사용되었다.

IF GML7 > 5.05 THEN above normal ELSE normal.

표 **3.2.** Final predictors of forecast models for each month

Month	Model(Threshold)	Final predictors	
		Forecasted Reanalysis data	Climate indices
Winter	REG	FCS10, FCP3, FBS1	GML7, AMO7
	Logistic REG(0.7)	INT	GML7
	Decision TREE(0.5)	*	GML7
	SVM	ALL	ALL
NOV	REG	*	AMO1
	Logistic REG(0.1)	FAWS3	GML1
	Decision TREE(0.5)	*	TNA1
	SVM	ALL	ALL
DEC	REG	INT, FCP3	NOI4
	Logistic REG(0.7)	FCP3	PDO4
	Decision TREE(0.5)	*	NOI4
	SVM	ALL	ALL
JAN	REG	INT, T, FBT11, FCS10	GML7, NOI8, WHW2
	Logistic REG(0.6)	FCS10	GML7
	Decision TREE(0.5)	FCS10	*
	SVM	ALL	ALL
FEB	REG	*	GML10
	Logistic REG(0.2)	FCT2	GML10
	Decision TREE(0.5)	*	GML10
	SVM	ALL	ALL

11월에 대하여 모형들을 적합한 결과, 의사결정나무는 normal로만 예측하고 있어 예측모형으로 사용할 수 없다. SVM은 100% 완벽한 정분류 결과를 보이고 있으나 모든 이차 잠재적 예측인자를 사용하고 있어 원인설명에 어려운 문제가 있으므로 예보자들이 꺼려하는 단점이 있다. 따라서 11월 황사출현일수에 대한 이 범주 예보는 다음의 식으로 구성되는 로지스틱 회귀모형을 예측모형으로 제안한다. 제안된 로지스틱 회귀모형에는 재분석자료 예측치에서 FAWS3와 광역규모 기후지수의 GML1이 적용되었다. 로지스틱 회귀모형에 사용된 문턱치는 0.1이다.

$$
\begin{aligned}
&X = 0.384*GML1 + 3.423*FAWS3\\
&EXP1 = EXP(X)\\
&P1 = EXP1/(1 + EXP1)\\
&P2 = 1 - P1\\
&\text{IF } P2 > 0.1 \text{ THEN above normal ELSE normal.}
\end{aligned}
$$

12월에 대하여 네 가지 모형들을 적합한 결과, 로지스틱 회귀모형과 의사결정나무의 결과가 유사하게 나타났으며, 그 중 두 종류의 예측인자가 고루 쓰이는 로지스틱 회귀모형을 예측모형으로 제안한다. 재분석자료 예측치인 FCP3과 광역규모 기후지수인 PDO4를 예측인자로 하는 로지스틱 회귀모형에 쓰인 문턱치는 0.7이다.

$$
\begin{aligned}
&X = 5.0678*FCP3 + 2.5055*PDO4\\
&EXP1 = EXP(X)\\
&P1 = EXP1/(1 + EXP1)\\
&P2 = 1 - P1
\end{aligned}
$$

IF P2 > 0.7 THEN above normal ELSE normal.

1월에 대하여 네 가지의 모형들을 적합한 결과, 네 모형 모두 우수한 예측결과를 보이고 있으며, 가장 예측력이 우수하다고 판단되는 중회귀모형을 예측모형으로 제안한다. 1월의 제안모형인 중회귀 모형에는 재분석자료 예측치인 FBT11과 FCS10, 광역규모 기후지수인 GML7, NOI8, WHW2가 적용되며 예측식은 다음과 같다.

$$X = 2.13562 - 0.01674*T + 0.37001*FBT11 + 0.60494*FCS10 - 0.08603*GML7 - 0.17304*NOI8 - 0.147010*WHW2$$

IF X > 0.55 THEN above normal ELSE normal.

2월에 대하여 모형들을 적합한 결과, 의사결정나무는 *normal*로만 예측하고 있어 예측모형으로 사용할 수 없고, 중회귀모형과 로지스틱 회귀모형은 FAR값이 50%이상 되며, SVM은 원인설명에 어려운 문제가 있어 예측모형 선택에 각각 문제가 있으나, 다음의 식으로 구성되는 중회귀모형이 가장 유용한 결과를 보이고 있어 예측모형으로 제안한다. 2월의 모형에 적용되는 변수는 광역규모 기후지수인 GML10이다.

$$X = 0.05290*GML10$$

IF X > 0.28 THEN above normal ELSE normal.

4. 요약 및 결론

겨울철 남한지역 황사출현일수 계절예측모형을 겨울철과 월별로 구분하여 다음 사항을 고려하여 개발하였다. 예보형태는 이 범주예보를 대상으로 하였다. 황사출현일수는 전국 28개 기상관서 관측자료의 공간적 평균을 사용하였다. 황사방출의 원인이 되는 황사발원지 지형조건을 고려하기 위하여 세 지역으로 구분하였으며, 황사발원지 기상상태에 따른 영향을 고려하기 위하여 네 기상요소(지상기온, 강우량, 강설량, 지상풍속)에 대한 NCEP 재분석자료 예측치와 황사운송에 영향을 주는 광역규모 기후지수를 잠재적 예측인자로 사용하였다. 예측모형 개발전략으로 세 방법을 고려하고, 네 모형(중회귀모형, 로지스틱 회귀모형, 의사결정나무, 지지벡터기계)을 적용하였다. 추정된 모형의 일반화 능력을 검증하기 위하여 교차검증을 수행하였다. 모형 평가는 2×2 분할표를 활용하여 정분류율, 탐지확률, 허위경고율, 검증결과를 종합적으로 판단하여 이루어졌다. 위의 사항을 고려한 황사 계절예측모형 개발을 통하여 다음의 결과들을 얻었다. 첫째, 황사발원지 기상요소 12종류에 대하여 각각 계절형 자기회귀-누적-이동평균모형을 적합한 결과, 모든 기상요소에서 주기 12가 검출되었다. 추정된 모형식의 통계적으로 유의하며, 예측치 생산에 유용한 것으로 판단된다. 둘째, 월별 황사출현일수에 대한 상관분석에서 상관성이 낮게 나타났다. 두 종류의 예측인자들(재분석자료 예측치, 광역규모 기후지수)과 황사출현일수 사이의 상관성 조사에서도 월에 따라 유의한 예측인자들이 다르게 선정되었다. 이는 월별로 구분하여 예측모형 개발의 필요성을 보여준다. 셋째, 겨울철 월별 황사 계절예측을 위한 예측모형들을 3절에서 제안하였다.

참고문헌

김연희, 김맹기, 이우섭 (2008). 한반도 기온 및 강수량 변동에 영향을 미치는 광역규모 기후지수들에 대한 고찰, *Atmosphere*, **18**, 86–95.

박순웅 (2002). Physical and chemical processes of Yellow Sand deflection, transport and transformation in East Asia, <한국과학재단 연구보고서>, (과제번호:R02–2001–000–00024–0), 1–63.

윤순창, 박경선 (1991). 등엔트로피 궤적에 의한 황사의 장거리 이동 경로 분석, <한국대기보전학회지>, **7**, 89–95.

전종갑, 예상옥, 권민호, 정용승 (2000). 한반도에서 관측된 1998년 4월 황사의 특성 및 장거리 수송패턴 분석, <한국기상학회지>, **36**, 405–416.

정용승, 김태균 (1991). 대기오염의 장거리 이동, 사례연구(황사,TSP,Sulphate의 발원지 추적), <한국대기보전학회지>, **7**, 197–202.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

Climate Diagnostic Center, *http://www.cpc.noaa.gov/data/indices*

Climate Prediction Center, *http://www.esrl.noaa.gov/psd/data/climateindices/list/*

Lim, J. Y. and Chun, Y. (2006). The characteristics of Asian dust events in Northeast Asia during the springtime from 1993 to 2004, *Global and Planetary Change*, **52**, 231–247.

NCEP, *http://nomad1.ncep.noaa.gov/pub/reanalysis-2/month/flx*

Park, S. and In, H. (2003). Parameterization of dust emission for the simulation of yellow sand (Asian dust) event observed on March 2002 in Korea, *Journal of Geophysical Research*, **108**, 4618.

Sohn, K. T., Cha, M. J., Chung, K. Y. and Song, S. J. (2008). Seasonal forecast of Asian dust over South Korea in spring season, *Journal of the Korean Data Analysis Society*, **10**, 2423–2433.

Sohn, K. T., Lee, J. H. and Cho, Y. S. (2009). Ternary forecast of Heavy snowfall in Honam area, Korea, *Advances in Atmospheric Sciences*, **26**, 327–332.

Sohn, K. T. and Park, S. M. (2008). Guidance on the choice of threshold for binary forecast modeling, *Advances in Atmospheric Sciences*, **25**, 83–88.

Tian, S. F., Inoue, M. and Du, M. (2007). Influence of dust storm frequency in Northern China on Fluctuations of Asian dust frequency observed in Japan, *SOLA*, **3**, 121–124.

Vapnik, V. (1996). *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

Zhang, B., Tsunekawa, A. and Tsubo, M. (2008). Contributions of sandy lands and stony deserts to long-distance dust emission in China and Mongolia during 2000–2006, *Global and Planetary Change*, **60**, 487–504.

Binary Forecast of Asian Dust Days over South Korea in the Winter Season

Keon-Tae Sohn[1] · Hyo-Jin Lee[2] · Seung-Bum Kim[3]

[1]Department of Statistics, Pusan National University
[2]Department of Statistics, Pusan National University
[3]National Institute of Meteorological Reseasrch

(Received March 2011; accepted April 2011)

Abstract

This study develops statistical models for the binary forecast of Asian dust days over South Korea in the winter season. For this study, we used three kinds of data; the first one is the observed Asian dust days for a period of 31 years (1980 to 2010) as target values, the second one is four meteorological factors(near surface temperature, precipitation, snowfall, ground wind speed) in the source regions of Asian dust based on the NCEP reanalysis data and the third one is the large-scale climate indices. Four kinds of statistical models(multiple regression models, logistic regression models, decision trees, and support vector machines) are applied and compared based on skill scores(hit rate, probability of detection and false alarm rate).

Keywords: Asian dust days, categorical forecast, NCEP reanalysis data, climate indices.

This research was carried out as a part of "Development of global forecast system and seasonal forecast models of Asian dust" supported by the National Institute of Meteorological Research in 2010.
[1]Corresponding author: Professor, Department of Statistics, Pusan National University, Busan 609-735, Korea. E-mail: ktsohn@pusan.ac.kr