기상 데이터를 이용한 산불 피해규모의 예측방법 연구

안승환*

<요약>

자연재해의 발생과 그 크기를 예측하는 것은 인명, 재산 피해를 미리 산정하고 그에 따른 대비책을 계획하는데 큰 도움이 된다. 본 연구는 기상 데이터를 이용하여 우리나라 산불의 특정 시점에서 피해규모를 예측하는 모형을 제안한다. 제안하는 모형은 산불의 발생 확률 예측 모형과, 산불이 발생했을 때 기대되는 피해규모를 예측하는 산불 피해규모 예측 모형으로 이루어져 있다.

확률 예측 모형의 적합은 정규화 로지스틱 회귀분석 모형, 단층 신경망 모형 및 최근 각광받고 있는 다층 신경망 모형을 이용하였다. 특히 다층 신경망 모형의 적합을 위해서 딥 러닝(Deep-learning) 알고리즘을 적용하고 성능을 비교하였다. 피해규모 예측 모형 역시 이와 유사한 방법을 적용하고 성능을 비교하였다.

본 연구에서는 2003년 1월부터 2016년 3월까지의 기간을 대상으로 연구를 진행하였다. 강수량, 상대습도, 온도, 풍속 등의 기상데이터를 이용하였고, 지역적인 요소를 고려하기 위해 우리나라를 20개의 지역으로 나누어 모형을 적합하였다. 그 결과 많은 지역에서 이 두 모형을 결합하여 상시적으로 특정지역의 산불피해 예측규모를 산정할 수 있었고, 이는 새로운 산불위험 지수로 사용할 수 있을 것이다.

주제어 : 산불 발생확률, 산불 피해규모, 딥 러닝(Deep-learning)

* 서울시립대학교 통계학과 학사과정

I. 서론

자연재해의 발생과 그 크기를 예측하는 것은 인명, 재산 피해를 미리 산정하고 그에 따른 대비책을 계획하는데 큰 도움이 된다. 우리나라는 국토의 대부분이 산림으로 이루어져 있기 때문에 많은 자연재해 중 산불은 특히 그 발생의 위험과 인명, 재산, 생태적 피해가 크다. 따라서 우리나라의 산불 발생을 예측하여 그 피해를 미리 산정하는 것은 매우 필요할 것이다. 산불의 발생은 대부분 인위적인 요인이 원인이 된다. 그러나 그것이 산림과 인명, 재산에 피해를 줄 정도의 발생에는 기상 요인이 많은 영향을 줄 것이다(이시영 등, 2004). 따라서 산불의 발생에 큰 영향을 주는 기상 데이터를 이용하여 우리나라 산불의 발생 확률과 피해규모를 분석하고 이들을 예측하는 모형을 연구한다. 최종적으로 앞의 두 모형의 적합을 통해 산불의 위험관리에 실용적인 기상 데이터가 주어졌을 때 산불 피해규모의 조건부 기댓값 모형을 적합하는 것을 목적으로 한다.

2003년 1월부터 2016년 3월까지 우리나라에서 발생한 산불에 대한 정보가 있는 산림청의 전국 산불발생 통계자료, 산불 발생일의 이전 날짜의 기상 데이터가 필요하므로 전국 산불발생 통계자료 시작일보다 1년 더 빠른 2002년 1월부터 2016년 3월까지의 기상청 기상 관측 자료를 연구에 이용하였다. 기상 관측 자료의 여러 기상 정보 중 캐나다 산불 예보 시스템의 구성 요소 중 하나인 FWI*(Forest Fire Weather Index) 지수를 계산하는데 필요한 강수량, 상대습도, 온도, 풍속 등이 산불의 발생에 중요한 영향을 미칠 것으로 판단하여 앞의 4가지 기상 정보를 선택하였다(Van Wagner, 1987). 연구에 사용될 기상 데이터 수집에는 R프로그램의 웹 크로울링(Web Crawling)방법을 이용하여 1시간 단위로 기상 데이터를 수집하였다. 또한 모형의 적합에서 지역적인 요소를 고려하기 위해 위, 경도를 이용해 격자모양으로 우리나라를 20개의 지역으로 나누어 이를 모형적합에 적용하였다.

핀란드에서는 기상 데이터와 실제 산불의 발생 자료에 기초하여 기상 데이터를 미세연료수분지수(Fine Fuel Moisture Code)로 변환하여 선형 회귀모형에 적용하여 지역에 따른 산불 발생 확률 모형의 적합에 대한 연구가 있다(Larjavaara et al, 2004). 우리나라에서도 산불의 발생 확률 모형에 대한 다양한 연구들이 진행되었다. 최관 등.(1996)은 대구, 경북 지역의 기상 자료를 로지스틱과 프로빗 확률 모형에 적용시켜 산불 발생 확률 모형을 적합하였다. 박홍석 등.(2009)은 강원도 지역의 기상 자료를 분석하고, 이를 이용하여 미세연료수분지수를 산출하여, 그 연중 분포와 로지스틱 모형에 적용시켜 확률 모형을 적합하고 그 적용성을 검토하는 연구를 하였다. 그리고 과거 산불발생 시계열자료와 전국 기상자료를 회귀분석모형에 적용하여 광역지별(도별) 산불 발생 확률 모형을 적합한 연구가 있다(이시영 등, 2004). 확

* FWI 지수는 프랑스 기상청과 캐나다 기상청에 의해 계산되는 산불 발생 위험의 추정값이다.

률 모형에 대한 기존의 연구들이 대부분 로지스틱 또는 회귀 분석방법을 사용하였다.

본 연구에서는 산불 발생 예측 확률 모형의 적합에 있어 새로운 분석 방법의 도입과 다양한 기상 요소의 고려를 통해 우수한 예측 성능을 가진 확률 모형의 적합을 목적으로 한다. 이를 위해 산불 발생일의 이전 14일 간의 강수량, 상대습도, 기온, 풍속 등의 다양한 기상 요인들을 이용하였다. 그리고 확률 예측 모형의 적합을 위해 기존의 확률 모형의 연구에서 많이 다루어지지 않은 분석 방법인 정규화 로지스틱 회귀분석 모형(Generalized Linear Model), 단층 신경망(Neural Network) 모형 및 다층 신경망(Multi-layered neural network) 모형 등을 이용하였고, 각각의 성능을 비교하여 가장 우수한 확률 모형을 제안한다. 확률 모형의 연구에서는 R프로그램의 'glmnet', 'nnet', 'h2o' 패키지에서 제공되는 함수들을 이용하였다.

SVM(Support Vector Machine)이나 유전 알고리즘(Genetic Algorithm) 등의 모형을 이용해 산불 발생 피해면적 예측 모형을 만든 기존의 연구들이 있다(Xiao, F., 2012; Xie, D. W. et al., 2014). Qu, X. L. et al.(2007)은 기상학적 요인들에 근거해 산불 피해면적을 예측하는 모형을 연구하였다. Safi, Y. et al.(2013)은 인공신경망(Artificial Neural Networks)를 이용하여 산불을 예측하는 모형을 제안하였다. 이병두 등.(2006)은 1970-2005년 동안의 산불 발생건수와 연소면적에 대해 시계열모형을 추정한 연구를 하였다. 그리고 FWI 지수 산출에 필요한 4가지 지수 데이터*, 지역변수, 4가지 기상 데이터와 실제 산불의 피해면적 데이터 등의 기상학적 데이터를 데이터 마이닝(Data Mining) 기법에 적용하여 산불이 발생하였을 때, 산불 피해면적을 예측하는 모형을 제시한 연구가 있다(P. Cortez, et al., 2007).

본 연구에서는 산불의 피해면적을 일정한 단계들로 나누어 설정하고 이를 산불 발생일의 이전 14일 간의 기상 데이터를 이용해 예측하는 모형을 제안한다. 즉, 본 연구에서 제안하는 모형은 산불이 발생하였을 때 기대되는 산불 피해규모를 예측하는 모형이다. 산불 피해규모 예측 모형의 적합을 위해 k-최근접 이웃 알고리즘(k-Nearest Neighbor Classification) 및 다층 신경망 모형 등을 이용하였다. R프로그램의 'kknn', 'h2o' 등의 패키지에서 제공되는 함수들을 이용하였다.

최종적으로 본 연구에서는 앞에서 제시한 두 모형을 적합한 모형인 기상 데이터를 이용해 산불 발생 피해규모를 예측하는 모형을 제안한다. 이 모형은 산불 발생일 이전의 14일 동안의 기상 데이터가 주어졌을 때 산불 피해규모의 조건부 기댓값을 구하는 모형이다. 따라서 특정일의 산불 피해규모의 기댓값을 알 수 있으므로 산불의 위험관리의 측면에서 그 효용성이 매우 크다고 생각된다.

본 논문은 우선 연구에 사용된 데이터를 소개하고, 연구 진행 과정을 소개한다. 연구 진행 과정은 데이터 가공, 산불발생 예측 확률 모형의 적합 과정, 산불이 발생

---

* FFMC(Fine Fuel Moisture Code), DMC(Duff Moisture Code), DC(Drought Code), ISI(Initial Spread Index) 지수

했을 때 기대되는 피해규모 예측 모형의 적합 과정 그리고 본 논문의 목적인 기상 데이터가 주어졌을 때 산불 피해규모 예측 모형의 적합 과정의 순서로 구성 되어있다. 그리고 결과 및 고찰, 결론으로 이 논문을 마무리한다.

II. 연구에 사용된 데이터 소개

1. 전국 산불발생 통계자료

우리나라에서 발생한 산불에 대한 데이터는 산림청이 제공하는 전국 산불발생 통계자료자료를 이용하였다. 전국 산불발생 통계자료자료에는 2003년부터 2016년 3월까지의 총 5476개의 산불 발생 관측치의 정보가 기록되어있다. 전국 산불발생 통계자료에 기록된 산불에 대한 정보는 발생일시, 진화종료시간, 발생장소, 발생원인, 피해면적(ha) 등이다. 본 연구에서는 발생일시, 발생장소, 피해면적(ha) 등의 3가지 산불 정보를 실제 산불 데이터로 사용하였다.

2. 기상청의 기상 관측 자료

현재 기상청의 기상 관측에 사용되는 총 기상대의 개수는 93개이다. 본 연구에서는 이들 93개 기상대에서 관측된 기상 관측 자료를 기상 데이터로 이용하였다. 연구에 사용된 전국 산불발생 통계자료가 2003년부터 기록되었고 그 이전의 기상 데이터가 필요하므로 2002년 1월 1일부터 데이터를 수집하였다. 그리고 전국 산불발생 통계자료의 마지막 기록이 2016년 3월 31일이므로 기상 관측 자료 또한 2016년 3월 31일까지의 데이터를 수집하였다.

데이터 수집은 매 날짜의 1시간 단위로 수집하였다. 기상 정보는 강수량(mm), 상대습도(%), 기온(℃), 풍속(m/s)들을 사용하였다. 데이터 수집은 기상청의 관측자료 제공 페이지를 대상으로 웹 크로울링(Web Crawling) 기법을 사용하여 수집하였다.

III. 연구 과정

1. 데이터 수집 및 가공

1) 기상 관측 자료의 수집 및 가공

기상 관측 자료는 1시간 단위로 수집하였으므로 각 일별로 강수량, 상대습도, 기온, 풍속에 대한 24개씩의 데이터가 있다. 즉, 각 일별로 96개의 기상 데이터가 있다. 강한 풍속과 낮은 습도는 산불의 대형화에 큰 영향을 미치는 요소이므로(박종길 등, 2014) 산불발생일마다 각 기상 정보의 24개 데이터에 대해 강수량의 평균값, 상

대습도의 최솟값, 기온의 최댓값, 풍속의 최댓값을 계산하여 평균 강수량, 최소 상대습도, 최고 기온, 최고 풍속을 연구에 사용할 기상 요인으로 선택하였다. 앞의 과정을 통해 기상 관측 자료를 각 일별로 평균 강수량, 최소 상대습도, 최고 기온, 최고 풍속의 4개의 기상 데이터가 있는 자료로 가공하였다.

2) 위, 경도를 이용한 데이터 가공

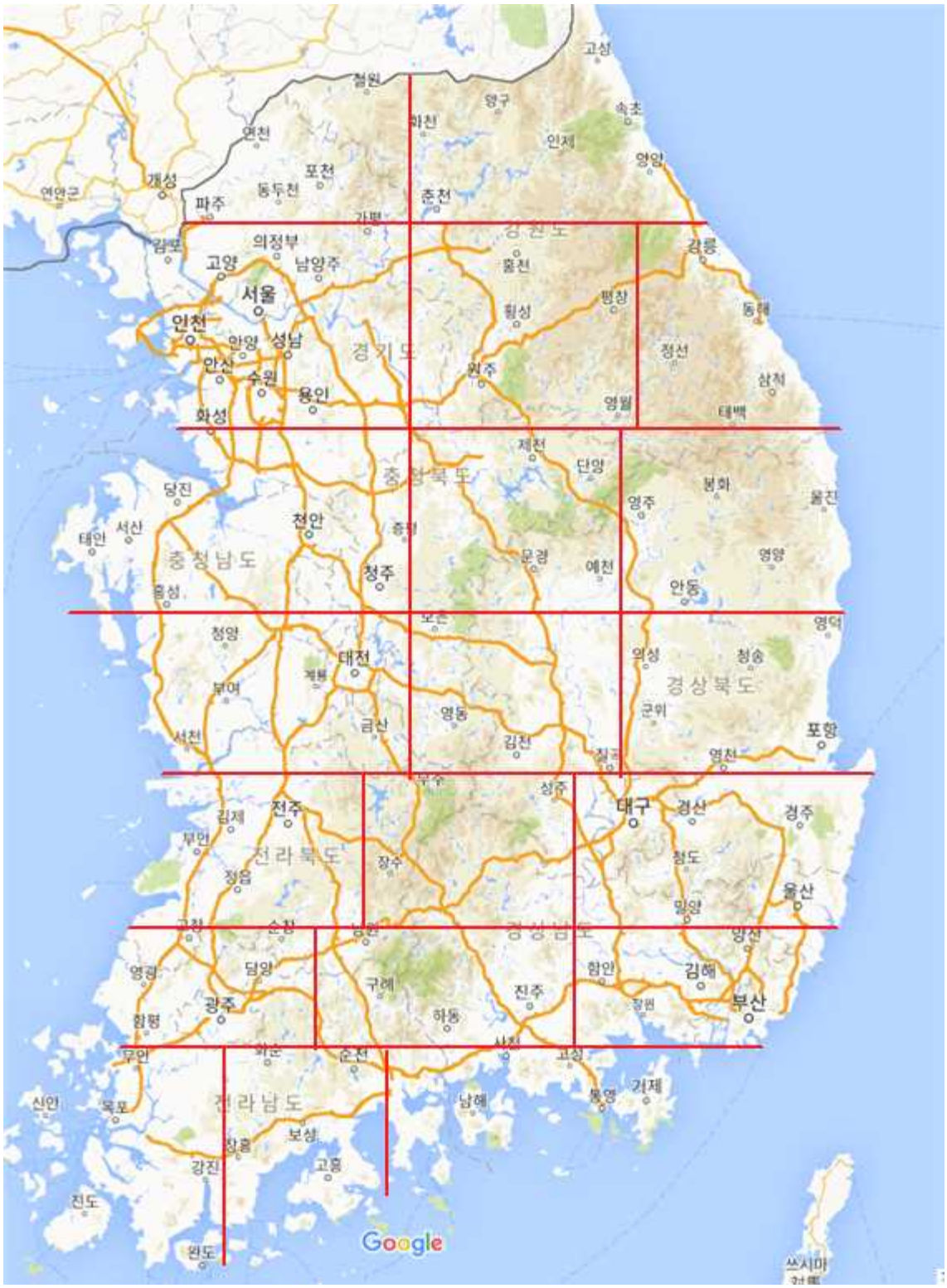
전국 산불발생 통계자료의 각 관측치의 산불발생 위치와 93개의 기상대의 위치를 구글 API에서 제공하는 geocode를 이용해 각 위치의 위, 경도 값을 구한다.

본 연구에서는 지역적인 요소를 고려하기 위해 전국을 몇 개의 지역으로 나누었다. 이때 행정단위를 기준으로 지역을 나누지 않고, 가까운 지역은 비슷한 지역적인 영향을 받을 것이므로 위, 경도에 따라 격자모양으로 20개 지역으로 나누었다(Cortez et al. 2007). 제주도는 산불 발생의 빈도가 매우 적어 지역의 구분과 연구 대상 지역에서 제외하였다. 20개 지역의 구분은 <그림 1>에 표시되어 있다. 그리고 20개 지역을 북서쪽 지역부터 차례대로 1~20의 번호를 붙였다.

다음으로 각 20개 지역마다 지역을 대표하는 새로운 기상 관측 자료를 가공한다. 우선 제주도에 위치한 4개의 기상대를 제외하고 나머지 기상대들의 위, 경도를 이용하여 앞에서 나눈 20개 지역의 위경도 기준에 따라 기상대마다 속하는 지역의 번호를 붙였다. 이 때 각 지역마다 3개에서 5개 정도의 기상대가 속하게 된다. 여러 개의 기상대 기상 관측 자료를 하나의 관측 자료로 통합하는 과정을 한다. 기상대 기상 관측 자료의 4가지 기상 데이터 값들에 대해 평균값, 최댓값, 최솟값을 선택하는 3가지 방법으로 한 지역에 속하는 기상대들의 기상 관측 자료를 각 일별로 4개의 기상 데이터가 있는 하나의 기상 관측 자료로 가공하여 통합한다. 모형의 적합을 위한 데이터의 사용에서 이전 14일의 기상 데이터를 사용하기로 하였다. 따라서 2003년 1월 1일부터 2016년 3월 31일까지 각 일마다 통합된 기상 관측 자료의 이전 14일의 기상 데이터, 즉 일별로 56개의 기상 데이터를 가지는 새로운 기상 관측 자료를 가공한다. 앞의 과정을 통해 새로운 20개의 가공된 기상대 기상 관측 자료가 생성되었다.

마지막으로 앞에서 구한 산불발생 위치의 위, 경도를 이용해 앞에서 나눈 20개 지역의 위경도 기준에 따라 전국 산불발생 통계자료의 각 관측치마다 산불 발생위치가 속하는 지역의 번호를 붙였다.

<그림 1> 본 연구에서 설정한 20개 지역의 구분을 나타낸 지도



자료: https://www.google.co.kr/maps/@36.4483731,127.674894,7.25z

2. 산불발생 확률 예측 모형의 적합

본 연구에서 만들고자 하는 확률 모형은 특정일 이전 14일의 기상 데이터가 주어졌을 때 산불 발생 확률을 예측하는 모형이다. 이 모형을 통해 특정일의 산불 발생 확률을 구하는 것이 확률 모형 적합의 목적이다.

1) 확률 모형 적합에 사용할 데이터 가공

우선 20개의 지역 중 전국 산불발생 통계자료의 각각의 관측치의 산불 발생 위치가 속하는 지역의 가공된 기상 관측 자료를 호출한다. 관측치의 발생일자와 일치하는 날짜의 호출된 가공된 기상 관측 자료의 관측치에 산불이 발생한 날이라는 의미로 1의 값을 부여한다. 나머지 일치하지 않는 날짜들에는 0의 값을 부여한다. 이 때 가공된 기상 관측 자료는 평균값, 최댓값, 최솟값 등을 선택하여 만든 3가지 자료 모두를 이용한다.

2) 산불 발생 확률 모형 적합

앞의 가공한 데이터를 바탕으로 확률 모형을 적합한다. 본 연구에서 고려하고 있는 확률 모형은 우도함수기반방법의 로지스틱 회귀분석방법이며, 모형의 복잡도를 조정하기 위해서 변수 선택에서 Lasso 벌점함수로 정규화(Generalized)하였다. 한편 모형의 비선형 관계를 반영하기 위해서 신경망 모형을 고려했으며, 모형의 복잡도를 효과적으로 조정하기 위해 다층 신경망 모형인 딥 러닝(Deep-learning) 알고리즘을 사용하였다. R프로그램의 glmnet 패키지, nnet 패키지, h2o 패키지 등에서 제공하는 함수를 이용하였다. 평균값, 최댓값, 최솟값을 선택한 20개 지역의 가공된 기상 관측 자료 모두에 대해 모형을 적합한다. 분류 성능을 알 수 있는 ROC curve*의 auc** 값을 통해 평균값, 최댓값, 최솟값의 선택에 따른 성능, 3가지 분석 방법에 따른 성능을 모두 비교한 뒤, 그 중 가장 성능이 우수한 모형을 최종 확률 모형으로 선택한다. 설명변수는 기상 관측 자료의 각 일별의 56개의 기상 데이터이고 반응변수는 산불 발생여부를 나타내는 0, 1의 값으로 이루어진 데이터이다. 모형의 train set은 가공된 기상 관측 자료의 2003년 1월 1일부터 2012년 12월 31일까지의 데이터이고 test set은 가공된 기상 관측 자료의 2013년 1월 1일부터 2016년 3월 31일까지의 데이터이다.

먼저 t시점의 산불의 확률 유무를 나타내는 베르누이 확률변수 Z_t(단, $Z_t=1$은 산불이 발생한 사건을 나타냄)와 그리고 $t-1$시점부터 s시점 이전까지의 기상 데이터를 나타내는 확률 벡터 $X_{t,s}=(X_{t-1},X_{t-2},\ldots,X_{t-s})$를 정의한다. X_t는 t시점의 평균 강수량, 최소 상대습도, 최고 기온, 최고 풍속의 4개의 기상 데이터를 나타낸다.

---

* ROC(Receiver Operating Characteristic) curve는 이진(binary) 분류기의 성능을 보여주는 그래프로, FPR(False Positive Rate) 또는 1-특이도(specificity)의 값을 x축, TPR(True Positive Rate) 또는 민감도(sensitivity)의 값을 y 축으로 가진다(특이도: 0인 경우에 대해 0으로 예측한 비율, 민감도: 1인 경우에 대해 1으로 예측한 비율)

** Area Under the Curve, ROC curve의 아래 면적을 나타내는 값으로 이진(binary) 분류기의 성능을 나타내는 지표이다. 값이 1에 가까울수록 이진 분류기의 성능이 좋다.

이때 s시점은 t시점의 14일 이전의 시점이 된다. 따라서 앞의 확률 벡터는 $\boldsymbol{X_{t,14}}=(X_{t-1},X_{t-2},\ldots,X_{t-14})$으로 다시 정의할 수 있다. 여기서부터 $\boldsymbol{X_{t,14}}$는 간단히 $\boldsymbol{X_t}$ 라고 표시하겠다. 본 연구에서 적합하고자 하는 확률 모형은 앞의 정의에 의해 $t-1$시점부터 14일 이전까지의 기상상태를 나타내는 확률 벡터 $\boldsymbol{X_t}$가 주어졌을 때 일별(t시점의) 산불 발생 확률 예측 모형이고 이는 t시점에서 산불이 발생할 확률 $P(Z_t=1|\boldsymbol{X_t})$으로 정의할 수 있다.

(1) 정규화 로지스틱 회귀분석 모형의 이용

첫 번째로 정규화 로지스틱 회귀분석 모형을 이용한 확률 모형의 적합이다. R프로그램의 glmnet 패키지를 이용하였고 적절한 모형선택을 위해 cv.glmnet 함수를 이용하여 조정계수를 선택하였다. 이 방법으로 선택된 모형의 회귀계수 β를 이용해 설명변수의 test set의 예측값을 구하였다.

정규화 로지스틱 회귀분석 모형을 이용해 적합한 t시점에서 산불이 발생할 확률 모형을 $P_{glm}(Z_t=1|\boldsymbol{X_t})$으로 나타낸다. 이제 $P_{glm}(Z_t=1|\boldsymbol{X_t})$를 설명변수의 test set의 예측값인 $\boldsymbol{X_t}'\beta$에 로지스틱 모형을 적용하여 모델링한다. 로지스틱 모형은 다음의 식 (1)과 같이 표현된다.

$$P_{glm}(Z_t=1|\boldsymbol{X_t})=\frac{1}{(1+\exp((\boldsymbol{X_t}'\beta)^{-1}))} \qquad (1)$$

이제 $P_{glm}(Z_t=1|\boldsymbol{X_t})$을 이용해 구한 설명변수의 test set의 산불 발생 확률값과 (실제 산불의 발생여부를 나타내는) 반응변수의 test set에 대해 ROC curve의 auc 값을 계산하여 모형의 성능을 구한다. 이는 AUC 패키지에서 제공하는 auc, roc 함수를 이용하였다. 평균값, 최댓값, 최솟값 중 하나를 선택한 기상 관측 자료의 20개 지역 각각에 대한 $P_{glm}(Z_t=1|\boldsymbol{X_t})$의 ROC curve의 auc 값은 <표 1>, <표 2>, <표 3>에 정리되어 있다.

(2) 단층 신경망(Neural Network) 모형 사용

두 번째로 단층 신경망 모형을 이용한 확률 모형 적합이다. 이는 nnet 패키지를 이용하였다. 우선 설명변수와 반응변수의 train set에 대해 신경망 모형의 내부 노드의 개수를 12개*로 잡고 신경망 모형의 복잡도를 조정하여** nnet 함수를 적용하였다. 얻어지는 모형은 t시점에서 산불이 발생할 예측 확률을 산출해 준다. 이 확률모형을 $P_{nnet}(Z_t=1|\boldsymbol{X_t})$이라 나타내고 다음의 식 (2)와 같이 표현된다.

$$P_{nnet}(Z_t=1|\boldsymbol{X_t})=f_{nnet}(\boldsymbol{X_t}) \qquad (2)$$

여기서 함수 f_{nnet}는 단층 신경망 모형 방법을 통해 추정한 함수이다.

다음 $P_{nnet}(Z_t=1|\boldsymbol{X_t})$을 이용해 설명변수의 test set의 확률값을 예측한다. 이는 predict 함수를 이용한다. 이제 $P_{nnet}(Z_t=1|\boldsymbol{X_t})$을 이용해 예측한 설명변수의 test

---

* nnet 함수에서는 size 옵션을 12로 하였다.

** nnet 함수에서는 decay 옵션을 0.05로하고 최대 반복수는 1000번으로 하였다.

set의 확률값과 반응변수의 test set에 대해 ROC curve의 auc 값을 계산하여 모형의 성능을 구한다. 평균값, 최댓값, 최솟값 중 하나를 선택한 기상 관측 자료의 20개 지역 각각에 대한 $P_{nnet}(Z_t = 1|X_{t,14})$의 ROC curve의 auc 값은 <표 1>, <표 2>, <표 3>에 정리되어 있다.

(3) 다층 신경망(Multi-layered neural network) 모형 사용

세 번째로 다층 신경망 모형을 사용한 확률 모형의 적합이다. 딥 러닝(Deep-learning) 알고리즘은 설명변수 내의 비선형적인 관계를 효과적으로 줄여주며, 이미지, 음성처리, 동영상 자료와 같은 특별한 데이터 분석에서 매우 뛰어난 성능을 가지고 있다는 것이 많은 연구를 통해 알려졌다. 따라서 본 연구에서도 설명변수의 시간적, 공간적 비선형관계를 고려, 정보를 효과적으로 축약하여 다층 신경망 모형을 적합하기 위해 딥 러닝 알고리즘을 이용하였다. h2o 패키지에서 제공하는 딥 러닝에 관한 함수를 이용하였다. 전체 반복수(epoch)는 10번, 활성함수(activation function)는 구간선형함수인 Rectified Linear Unit(ReLU), 내부층의 개수는 3개이고 하나의 층에서 노드 개수는 200개, 그리고 추정계수의 절반을 버리는* 옵션을 사용하였다. 앞의 과정을 통해 t시점에서 산불이 발생할 확률 예측 모형을 만든다. 이 확률 모형을 $P_{dl}(Z_t = 1|\boldsymbol{X_t})$이라 나타내고 다음의 식 (3)와 같이 표현된다.

$$P_{dl}(Z_t = 1|\boldsymbol{X_t}) = f_{dl}(\boldsymbol{X_t}) \qquad (3)$$

여기서 함수 f_{dl}는 딥 러닝(Deep-learning) 분석 방법을 통해 추정한 함수이다.

다음 $P_{dl}(Z_t = 1|\boldsymbol{X_t})$을 이용해 설명변수의 test set에 대해 확률값을 예측한다. 이는 h2o.predict 함수를 이용한다. 이제 $P_{dl}(Z_t = 1|\boldsymbol{X_t})$을 이용해 예측한 설명변수의 test set의 확률값과 반응변수의 test set에 대해 ROC curve의 auc 값을 계산하여 모형의 성능을 구한다. 평균값, 최댓값, 최솟값 중 하나를 선택한 기상 관측 자료의 20개 지역 각각에 대한 $P_{dl}(Z_t = 1|\boldsymbol{X_t})$의 ROC curve의 auc 값은 <표 1>, <표 2>, <표 3>에 정리되어 있다.

* 함수에서는 RectifierWtihDropout 을 TRUE로 설정하였다.

(4) 산불 발생 확률 모형의 선택

<표 1> 평균값을 선택해 가공한 20개 지역의 기상 관측 자료를 이용했을 때 3가지 확률 모형들의 ROC curve의 auc 값

	확률 모형		
지역 번호	$P_{glm}(Z_t = 1\|\boldsymbol{X_t})$	$P_{nnet}(Z_t = 1\|\boldsymbol{X_t})$	$P_{dl}(Z_t = 1\|\boldsymbol{X_t})$
1	0.71816191	0.646733339	0.736149333
2	0.693100108	0.61249242	0.681467714
3	0.727271659	0.726872065	0.718515863
4	0.710271523	0.59566213	0.723458989
5	0.687893338	0.626746228	0.729616335
6	0.754204584	0.658302371	0.755371278
7	0.682958329	0.594377798	0.704583286
8	0.692265227	0.627590759	0.694917492
9	0.774439746	0.636918605	0.761839323
10	0.683269243	0.685670105	0.75454231
11	0.718598862	0.628520626	0.689378853
12	0.759993692	0.700149807	0.756866146
13	0.735024155	0.580591787	0.728478261
14	0.728568951	0.674942979	0.727328505
15	0.774241466	0.771088812	0.78931732
16	0.698599034	0.507669082	0.685048309
17	0.697293729	0.68340234	0.694665467
18	0.750194351	0.707558089	0.771823443
19	0.765427288	0.715454796	0.758389877
20	0.8187751	0.76507817	0.778650316

<표 2> 최댓값을 선택해 가공한 20개 지역의 기상 관측 자료를 이용했을 때 3가지 확률 모형들의 ROC curve의 auc 값

	확률 모형		
지역 번호	$P_{glm}(Z_t = 1\|\boldsymbol{X_t})$	$P_{nnet}(Z_t = 1\|\boldsymbol{X_t})$	$P_{dl}(Z_t = 1\|\boldsymbol{X_t})$
1	0.705210029	0.68105114	0.714555057
2	0.679750813	0.584738488	0.701747616
3	0.702602646	0.701829903	0.68492064
4	0.681168638	0.620524672	0.692151126
5	0.687085295	0.606381987	0.73716843
6	0.741532497	0.646931596	0.734819521
7	0.697494547	0.59782172	0.694208472
8	0.68120012	0.651989199	0.690813081
9	0.759534884	0.692859408	0.736384778
10	0.65555759	0.599045759	0.747604224
11	0.721028924	0.670791844	0.689224751
12	0.703881837	0.620331678	0.757628321
13	0.705772947	0.611328502	0.738067633
14	0.5	0.674055993	0.709828872
15	0.765628951	0.747281922	0.775
16	0.7	0.637403382	0.707922705
17	0.708910891	0.61990399	0.716327633
18	0.739310702	0.744441565	0.726008465
19	0.745759215	0.621699065	0.759238034
20	0.813683305	0.704012478	0.764773379

<표 3> 최솟값을 선택해 가공한 20개 지역의 기상 관측 자료를 이용했을 때 3가지 확률 모형들의 ROC curve의 auc 값

	확률 모형		
지역 번호	$P_{glm}(Z_t = 1\|\boldsymbol{X_t})$	$P_{nnet}(Z_t = 1\|\boldsymbol{X_t})$	$P_{dl}(Z_t = 1\|\boldsymbol{X_t})$
1	0.680828639	0.555701287	0.708840303
2	0.553916974	0.506296614	0.583958794
3	0.711893191	0.714508177	0.702755431
4	0.690808351	0.585767997	0.692546891
5	0.603491679	0.599451463	0.613506752
6	0.777125626	0.63815447	0.783120636
7	0.678208587	0.512261795	0.678366433
8	0.67669967	0.57780378	0.684092409
9	0.610824524	0.520264271	0.645158562
10	0.706748866	0.607916743	0.740259212
11	0.636024182	0.583677098	0.654789
12	0.784514705	0.675655076	0.73810087
13	0.601123188	0.586256039	0.675410628
14	0.688114388	0.561575501	0.697664493
15	0.646523388	0.510698483	0.672819216
16	0.624661836	0.577246377	0.625193237
17	0.668922892	0.533363336	0.640912091
18	0.601122916	0.401382051	0.601934871
19	0.744017055	0.610466716	0.738515496
20	0.813683305	0.746647303	0.776391279

위의 <표 1>, <표 2>, <표 3>을 보면 평균값, 최댓값, 최솟값을 선택해 가공한 20개 지역의 기상 관측 자료를 이용했을 때 3가지 확률 모형들의 ROC curve의 auc 값이 0.7보다 큰 확률 모형의 개수는 각각 34, 31, 15개 이다. 따라서 ROC curve의 auc 값이 0.7보다 큰 확률 모형이 가장 많은 경우인 평균값을 선택해 가공한 20개 지역의 기상 관측 자료를 사용했을 때 확률 모형들의 성능이 다른 경우에 비해 우수하다고 할 수 있다. 또한 평균값을 선택해 가공한 20개 지역의 기상 관측 자료를 사용한 경우에서 20개 지역에 대한 확률 모형 중 딥 러닝 방법을 이용해 만든 확률 모형($P_{dl}(Z_t = 1|\boldsymbol{X_t})$)들이 ROC curve의 auc 값이 0.7보다 큰 확률 모형의 개수가 15개로 가장 많다(정규화 로지스틱 회귀분석 모형($P_{glm}(Z_t = 1|\boldsymbol{X_t})$): 13개, 단층 신경망 모형($P_{nnet}(Z_t = 1|\boldsymbol{X_t})$): 6개). 따라서 산불 발생 확률 예측 모형 중 평균값

을 선택해 가공한 20개 지역의 기상 관측 자료와 딥 러닝 알고리즘을 이용해 적합한 다층 신경망 모형이 20개 지역에 대해서 가장 우수한 성능을 가지는 확률 모형이다.

따라서 최종적으로 본 연구에서는 산불 발생 확률 예측 모형으로 평균값을 선택해 가공한 20개 지역의 기상 관측 자료를 이용한 확률 모형인 $P_{dl}(Z_t = 1|\boldsymbol{X_t})$를 제안하고 이를 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$라고 나타낸다. 20개 지역에 대한 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$의 ROC curve 그래프는 <부록 1>의 <그림 5>~<그림 24>에 있다.

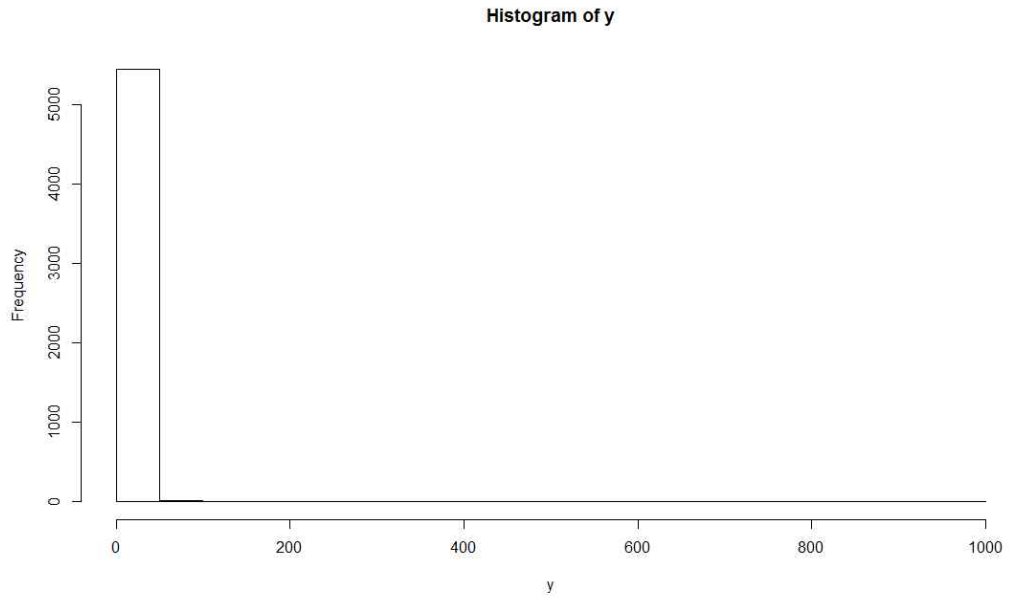
3. 산불이 발생했을 때 기대되는 피해규모의 예측 모형

본 연구에서 만들고자 하는 산불이 발생했을 때 기대되는 피해규모의 예측 모형은 산불이 발생하였다는 조건과 이전 14일의 기상 데이터가 주어졌을 때 기대되는 산불 피해규모를 예측하는 모형이다.

1) 산불 피해규모 예측 모형 적합에 사용할 데이터 가공

우선 20개의 지역 중 전국 산불발생 통계자료의 각각의 관측치의 산불 발생 위치가 속하는 지역의 기상 관측 자료를 호출한다. 다음 전국 산불발생 통계자료의 관측치의 날짜와 일치하는 날짜의 가공된 기상 관측 자료의 날씨 데이터를 전국 산불발생 통계자료의 관측치의 피해면적 정보와 합친다. 앞과 같은 과정을 통해 전국 산불발생 통계자료의 관측치가 속하는 지역의 관측치의 날짜의 14일 이전 동안의 날씨 데이터를 설명변수로 가지는 가공된 전국 산불발생 통계자료를 만든다.

<그림 2> 산불 피해면적 데이터의 히스토그램



위의 <그림 2>은 가공된 전국 산불발생 통계자료의 산불 피해면적 데이터의 히스토그램을 나타낸 것이다. 히스토그램에서 알 수 있듯이 산불 피해면적 데이터는 연속적인 값들이 아닌 0.01부터 973까지의 일부 값들에 집중되어 반복적으로 나타난다. 따라서 산불 피해면적의 값을 예측하는 것이 아닌 일정한 기준에 따라 피해규모 단계를 만들어 이를 기댓값으로 예측하는 방법을 사용하였다. 즉, 반응변수가 1에서 7까지에 해당하는 산불 피해규모 단계로 구성된다. 산불 피해규모 단계는 다음의 <표 4>와 같다.

<표 4> 산불 피해규모 단계 표

단계	산불 피해면적 크기(ha)
1	0이상 0.1미만
2	0.1이상 0.3미만
3	0.3이상 0.7미만
4	0.7이상 1미만
5	1이상 2미만
6	2이상 3미만
7	3이상

<표 5> 가공된 전국 산불발생 통계자료의 산불 피해면적 데이터에서 각 산불 피해규모 단계의 빈도수 표

단계	빈도수
1	1721
2	1770
3	1069
4	195
5	377
6	113
7	231
총합	5476

<표 4>와 같이 피해규모 단계를 설정한 이유는 산불 피해크기가 1보다 작은 값의 빈도가 매우 높고 3이상의 큰 값은 매우 드물게 나타나기 때문에 빈도가 높은 산불 피해크기 값들은 더 세분화해서 단계를 설정하였고 빈도가 낮은 산불 피해크기 값들은 하나로 묶어 단계를 설정하였다. 위의 <표 5>은 가공된 전국 산불발생 통계자료의 산불 피해면적 데이터에서 각 산불 피해규모 단계들의 빈도수를 나타내었다.

2) 산불이 발생했을 때 기대되는 피해규모 예측 모형의 적합

앞의 가공한 데이터를 바탕으로 산불 피해규모 예측 모형을 적합한다. 본 연구에서 고려하고 있는 피해규모 예측 모형은 k-최근접 이웃 알고리즘(k-Nearest Neighbor Classification)을 고려했으며, 모형의 비선형 관계를 반영하고 복잡도를 효과적으로 조정하기 위해 다층 신경망 모형인 딥 러닝(Deep-learning) 알고리즘을 사용하였다. R프로그램의 kknn 패키지, h2o 패키지 등에서 제공하는 함수를 이용하였다. 평균값, 최댓값, 최솟값을 선택해 가공한 20개 지역의 기상 관측 자료 모두를 이용해 가공된 전국 산불발생 통계자료를 만들고 이를 이용해 모형을 적합한다.

산불 피해규모 예측 모형의 성능을 알 수 있는 반응변수의 test set과 모형의 설명변수의 test set의 피해규모 예측값과의 일치도를 이용해 평균값, 최댓값, 최솟값의 선택에 따른 성능, 2가지 분석 방법에 따른 성능을 모두 비교한 뒤, 그 중 가장 성능이 우수한 모형을 산불 피해크기 예측 모형으로 선택한다. 설명변수는 기상 관측 자료 관측치의 56개의 기상 데이터이고 반응변수는 가공된 전국 산불발생 통계자료의 산불 피해규모 데이터이다. 모형의 train set은 가공된 전국 산불발생 통계자료의 각 산불 피해규모 단계들에 속하는 관측치들의 75%를 선택하였고, 나머지 관측치들을 test set으로 선택하였다.

우선 t시점에서 산불이 발생했을 때 산불의 피해규모를 나타내는 확률 변수 Y_t를 정의한다. 본 연구에서 적합하고자 하는 산불이 발생했을 때 기대되는 피해규모 예측 모형은 확률 벡터 $\boldsymbol{X_t}$가 주어졌을 때 산불 피해규모 예측 모형이고 이는 Y_t의 조건부 기댓값으로 $E(Y_t|Z_t=1,\boldsymbol{X_t})$와 같이 나타낸다. 이는 식 (4)와 같다.

$$E(Y_t|Z_t=1,\boldsymbol{X_t})=f(\boldsymbol{X_t}) \qquad (4)$$

여기서 함수 f는 전국 산불발생 통계자료의 train set과 test set의 각 관측치의 $\boldsymbol{X_t}$에 따라 산불 피해크기의 기댓값을 1에서 7의 단계로 분류하여 예측해주는 함수이다. 함수 f는 k-최근접 이웃 알고리즘(k-Nearest Neighbor Classification) 및 딥 러닝 알고리즘 등의 방법을 이용하여 추정한다.

(1) k-최근접 이웃 알고리즘(k-Nearest Neighbor Classification) 사용

첫 번째로 k-최근접 이웃 알고리즘을 이용한 산불이 발생하였을 때 기대되는 피해면적 예측 모형 적합이다. 이는 R프로그램의 kknn 패키지를 이용하였다. 설명변수와 반응변수의 train set, test set 모두에 대해 kknn 함수를 적용하여 t시점에서

산불이 발생하였을 때 기대되는 피해규모 예측 모형과 산불 피해규모 예측값을 구한다. 이 피해규모 예측 모형을 $E_{knn}(Y_t|Z_t=1,\boldsymbol{X_t})$이라 나타내고 다음의 식 (5)와 같다.

$$E_{knn}(Y_t|Z_t=1,\boldsymbol{X_t})=f_{knn}(\boldsymbol{X_t}) \quad (5)$$

여기서 함수 f_{knn}는 k-최근접 이웃 알고리즘을 통해 추정한 함수이다.

다음 $E_{knn}(Y_t|Z_t=1,\boldsymbol{X_t})$의 성능을 알아보기 위해 설명변수의 test set의 예측값과 반응변수의 test set과의 일치도를 검사한다. 이 때 모형의 적합에는 전국 산불발생 통계자료는 평균값, 최댓값, 최솟값을 선택해 가공한 3가지 모두가 이용되었다. 이 모형의 일치도는 <표 6>에 정리되어 있다.

(2) 다층 신경망(Multi-layered neural network) 모형 사용

두 번째로 다층 신경망 모형을 사용한 확률 모형의 적합이다. 본 연구에서는 설명변수의 시간적, 공간적 비선형관계를 고려, 정보를 효과적으로 축약하여 다층 신경망 모형을 적합하기 위해 딥 러닝 알고리즘을 이용하였다. h2o 패키지에서 제공하는 딥 러닝에 관한 함수를 이용하였다. 전체 반복수(epoch)는 10번, 활성함수(activation function)는 구간선형함수인 Rectified Linear Unit(ReLU), 내부층의 개수는 2개이고 하나의 층에서 노드 개수는 200개, 그리고 추정계수의 절반을 버리는* 옵션을 사용하였다. 앞의 과정을 통해 t시점에서 산불이 발생하였을 때 기대되는 피해규모 예측 모형을 만든다. 이 모형을 $E_{dl}(Y_t|Z_t=1,\boldsymbol{X_t})$이라 나타내고 다음의 식 (7)와 같이 표현된다.

$$E_{dl}(Y_t|Z_t=1,\boldsymbol{X_t})=f_{dl}(\boldsymbol{X_t}) \quad (7)$$

여기서 함수 f_{dl}는 딥 러닝(Deep-learning) 분석 방법을 통해 추정한 함수이다.

다음 $E_{dl}(Y_t|Z_t=1,\boldsymbol{X_t})$의 성능 알아보기 위해 설명변수의 test set의 예측값과 반응변수의 test set과의 일치도를 검사한다. 설명변수의 test set의 예측값은 $E_{dl}(Y_t|Z_t=1,\boldsymbol{X_t})$와 설명변수의 test set에 h2o.predict 함수를 적용시켜 구한다. 이 때 모형의 적합에는 전국 산불발생 통계자료는 평균값, 최댓값, 최솟값을 선택해 가공한 3가지 모두가 이용되었다. 이 모형의 일치도는 <표 6>에 정리되어 있다.

* 함수에서는 RectifierWtihDropout 을 TRUE로 설정하였다.

(3) 산불이 발생했을 때 기대되는 피해면적의 예측 모형의 선택

<표 6> t시점에서 산불이 발생했을 때 기대되는 피해규모 예측 모형의 설명변수의 test set의 예측값과 반응변수의 test set과의 일치도

가공된 기상 관측 자료의 종류	$E_{knn}(Y_t\|Z_t=1,\boldsymbol{X_t})$	$E_{dl}(Y_t\|Z_t=1,\boldsymbol{X_t})$
평균값 선택	0.2206846	0.3190095
최댓값 선택	0.2301529	0.3182811
최솟값 선택	0.2600146	0.3182811

6개의 t시점에서 산불이 발생했을 때 기대되는 피해규모 예측 모형들의 일치도 중 평균값을 선택해 가공한 전국 산불발생 통계자료와 딥 러닝 알고리즘을 이용해 적합한 모형의 일치도 값이 가장 큰 것을 볼 수 있다. 따라서 본 연구의 t시점에서 산불이 발생하였을 때 기대되는 피해규모 예측 모형으로 평균값을 선택해 가공한 전국 산불발생 통계자료와 딥 러닝 알고리즘을 이용해 적합한 모형을 최종적으로 선택하고 이를 $E_{dl,mean}(Y_t|Z_t=1,\boldsymbol{X_t})$으로 정의한다.

4. 기상 데이터가 주어졌을 때 기대되는 산불 피해규모 예측 모형

본 연구의 최종 목적이 되는 모형은 특정일 이전 14일의 기상 데이터가 주어졌을 때 산불 피해규모를 예측하는 모형이다. 이는 앞에서 구한 20개 지역의 산불 발생의 조건부 확률 모형과 산불 피해규모의 조건부 기댓값 모형의 적합으로 표현할 수 있다.

앞에서 정의한 모형의 추정량들과 조건부 기댓값의 정의를 이용하면, $\boldsymbol{X_t}$가 주어져 있을 때 Y_t의 기댓값에 대한 모형으로 다음의 식 (8)과 같은 특정 시점 t에서 기대되는 산불 피해규모 예측 모형을 나타낼 수 있다. 이 식은 본 연구에서 제안하고자 하는 기상 데이터가 주어졌을 때 기대되는 산불 피해규모 예측 모형과 같다.

$$E(Y_t|\boldsymbol{X_t}) = E(Y_t|Z_t=1,\boldsymbol{X_t})P(Z_t=1|\boldsymbol{X_t}) \quad (8)$$

IV. 결과 및 고찰

1. 산불 발생 확률 모형의 결과 및 고찰

본 연구에서 최종적으로 선택한 산불 발생 확률 모형인 $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$은 20개의 대부분의 지역에서 좋은 성능을 보여주었으나, 몇 개의 지역에서는 다른 지역에

비해 그 성능이 떨어졌다. 모형의 성능이 다소 떨어지는 지역은 2, 8, 11, 16, 17번 지역이며, 이 지역들은 대부분 우리나라의 동쪽 지역이며 산림이 많은 지역들이다. 이러한 지역들에서 모형의 성능이 약간 떨어지는 이유는 산림이 많아 산불의 발생에 있어 기상적 요인뿐만 아니라 본 연구에서 모형의 적합에서 고려하지 않은 지형, 산림 조건이 영향을 미치기 때문이라고 고려된다. 한편으로는 산림이 적은 나머지 지역에 대해서는 지형, 산림 조건보다 기상적 요인이 산불의 발생에 주요한 영향을 미치므로 본 연구에서 제안한 모형의 성능이 높았다.

다음은 $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$은 해석이 매우 어려우므로 보다 해석이 쉬운 정규화 로지스틱 회귀분석 모형에 대한 해석이다. 이 때 ROC curve의 auc 값이 가장 큰 모형인 평균값을 선택해 가공한 기상 관측 모형을 이용해 적합한 9번 지역의 $P_{glm}(Z_t=1|\boldsymbol{X_t})$을 예시로 들었고, 이를 $P_{glm,mean}(Z_t=1|\boldsymbol{X_t})$이라 정의한다. 다음의 <표 7>은 $P_{glm,mean}(Z_t=1|\boldsymbol{X_t})$의 회귀계수($\beta_1$~$\beta_{56}$)를 정리한 표이다.

<표 7> $P_{glm,mean}(Z_t=1|\boldsymbol{X_t})$의 회귀계수들을 정리한 표

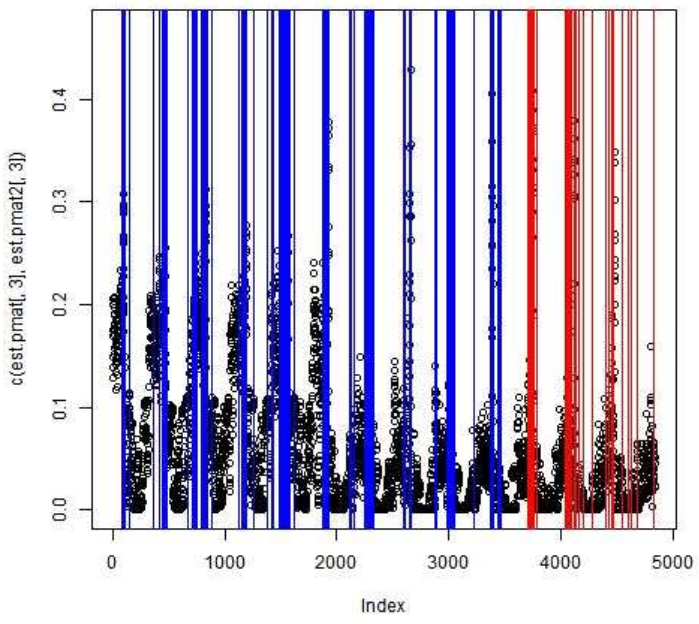
β_1	-0.14427	β_2	-0.02596	β_3	0.01578	β_4	0.02687
β_5	-0.08328	β_6	-0.00978	β_7	0	β_8	0.050265
β_9	-0.08624	β_{10}	0	β_{11}	0.035164	β_{12}	0
β_{13}	-0.03182	β_{14}	-0.00515	β_{15}	0.021876	β_{16}	-0.02654
β_{17}	-0.0335	β_{18}	0	β_{19}	0	β_{20}	-0.05456
β_{21}	-0.02358	β_{22}	0	β_{23}	0	β_{24}	0
β_{25}	-0.00097	β_{26}	0.004917	β_{27}	-0.00128	β_{28}	0
β_{29}	-0.01599	β_{30}	0	β_{31}	0	β_{32}	0.040722
β_{33}	0	β_{34}	0	β_{35}	0	β_{36}	0.117373
β_{37}	-0.01471	β_{38}	0	β_{39}	0	β_{40}	0
β_{41}	-0.00535	β_{42}	0	β_{43}	-0.01548	β_{44}	0
β_{45}	-0.01149	β_{46}	0	β_{47}	0	β_{48}	0.079836
β_{49}	-0.00061	β_{50}	0	β_{51}	-0.00885	β_{52}	0.217998
β_{53}	-0.02088	β_{54}	0.002758	β_{55}	-0.14081	β_{56}	0.04148

<표 7>의 1, 2열에 해당하는 회귀계수들은 평균 강수량 변수들의 회귀계수이다. 이 회귀계수들의 집합을 β_r이라 정의한다. <표 7>의 3, 4열에 해당하는 회귀계수들은 평균 상대습도 변수들의 회귀계수이다. 이 회귀계수들의 집합을 β_h이라 정의한다. <표 7>의 5, 6열에 해당하는 회귀계수들은 최고 기온 변수들의 회귀계수이다. 이 회귀계수들의 집합을 β_t이라 정의한다. <표 7>의 7, 8열에 해당하는 회귀계수들은 최대 풍속 변수들의 회귀계수이다. 이 회귀계수들의 집합을 β_w이라 정의한다. 회

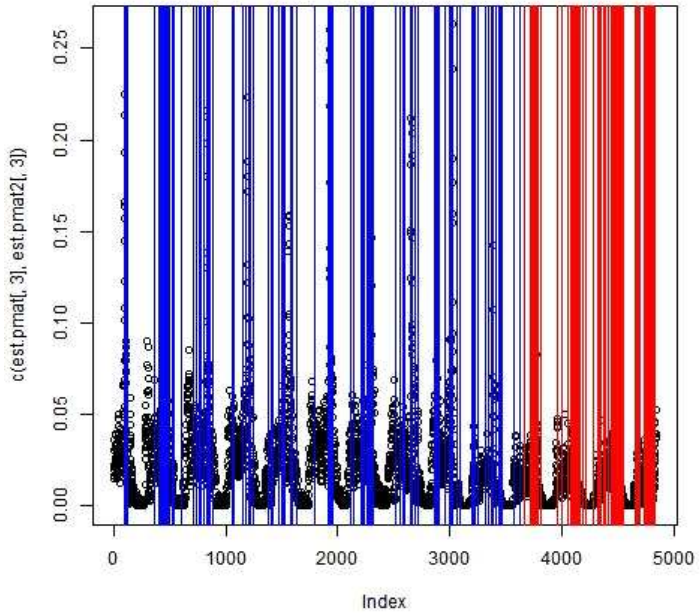
귀계수의 집합 β_r, β_h, β_t, β_w 각각에 속하는 회귀계수들에 대해서 회귀계수의 아래 첨자가 작을수록 특정일로부터 가까운 시점의 기상 데이터임을 의미한다.

β_r에 속하는 회귀계수들이 모두 음수이므로 특정일 이전의 기간 동안 비가 적게 내릴수록 산불 발생의 확률이 높아진다. 또한 β_r는 다른 회귀계수 집합과 비교했을 때 값이 0인 회귀계수를 1개 밖에 갖지 않으므로 특정일 이전의 평균 강수량 데이터는 4가지 기상 데이터 중 산불 발생 확률에 가장 많은 영향을 미친다고 할 수 있다. β_h는 다른 회귀계수 집합에 비해 0인 회귀계수를 가장 많이 가지고 있다. 따라서 특정일 이전의 평균 상대습도 데이터는 다른 기상 데이터에 비해 산불 발생 확률에 적은 영향을 미친다고 할 수 있다. 그리고 특정일로부터 가까운 시점의 평균 상대습도가 낮을수록 산불 발생확률이 높아지고, 먼 시점의 평균 상대습도는 거의 영향을 미치지 않는다. 이는 평균 상대습도가 낮으면 날씨가 건조해지므로 산불 발생 확률이 높아지는 것과 관련이 있다. β_t에 속하는 회귀계수들을 통해 특정일로부터 가까운 시점의 최고기온은 높고, 특정일로부터 먼 시점의 최고기온은 낮을수록 산불 발생 확률이 높아지는 것을 알 수 있다. 마지막으로 β_w에 속하는 회귀계수들을 통해 특정일 이전 대부분의 시점에서 최고 풍속이 클수록 산불 발생 확률이 높아지는 것을 알 수 있다.

<그림 3> $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$의 15번 지역에 대한 확률 좌표 그래프



<그림 4> $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$의 2번 지역에 대한 확률 좌표 그래프



위의 <그림 3>, <그림 4>의 그래프는 각각 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$의 15번, 2번 지역에 대한 실제 산불발생의 예측 정도를 보여주는 그래프이다. 위의 그래프의 설명은 다음과 같다. 그래프의 x축은 1부터 2003년 1월 1일을 나타내고 차례로 전체 연구 대상 기간을 나타낸다. 그래프의 y축은 산불 발생 확률을 나타낸다. 그래프에 찍혀 있는 좌표들은 설명변수의 train set과 test set 전체에 대해 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$을 이용해 예측한 산불 발생 확률을 나타낸다. 즉, 이 좌표들의 y값은 각 좌표에 해당하는 관측치의 날짜, 즉 x축의 값이 대응되는 날짜의 산불 발생 확률을 나타낸다. 그리고 그래프위에 그려진 세로선은 실제 산불 발생 일에 대응되는 x축의 값에 그린 것이다. 특히 푸른색 세로선은 실제 산불 발생일 중 train set에 속하는 날짜들을 그 해당하는 x값에 그린 것이고, 붉은색 세로선은 실제 산불 발생일 중 test set에 속하는 날짜들을 그 해당하는 x값에 그린 것이다. 따라서 이 그래프들은 컬러로 보기를 권장한다. 이 그래프를 확률 좌표 그래프라고 정의한다.

15번 지역과 2번 지역을 예시로 든 이유는 20개의 전체 지역 중 15번 지역의 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$의 ROC curve의 auc 값이 가장 크고 2번 지역의 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$의 ROC curve의 auc 값이 가장 작아 두 지역의 확률 좌표 그래프를 비교하기 위함이다.

우선 15번 지역의 확률 좌표 그래프(<그림 3>)를 보자. y값이 큰 좌표들의 x값에 세로선이 겹쳐져서 그려져 있는 것을 볼 수 있다. 특히 설명변수의 test set을 이용한 것에 해당되는 좌표(x값이 3654부터 4839까지의 좌표)들에 대해 큰 y값, 즉 높은 확률값을 가지는 좌표들에 x값에 붉은색 세로선이 거의 겹쳐져서 그려져 있고,

y값(확률값)이 작은 좌표의 x값에는 붉은색 세로선이 거의 겹쳐져 있지 않다. 즉, $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$를 이용해 예측한 산불 발생의 확률이 높을 때 실제로 산불이 발생하였으므로 확률 모형의 이진 분류의 성능이 우수하고 따라서 $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$의 ROC curve의 auc 값이 높게 나타나게 되고 $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$을 이용해 상당히 정확한 산불 발생 확률을 예측할 수 있다.

다음은 2번 지역의 확률 좌표 그래프(<그림 4>)를 보자. 2번 지역에 대한 $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$의 ROC curve의 auc 값은 다른 지역에 비해 낮게 나왔다. 이 이유를 그래프를 통해 알 수 있다. 설명변수의 test set을 이용한 것에 해당되는 좌표(x값이 3654부터 4839까지의 좌표)들에 대해 특별히 높은 y값(확률값)을 가지는 좌표가 없지만 같은 기간 동안 붉은색 세로선이 매우 밀도 높게 그려져 있는 것으로 보아 실제 산불은 높은 빈도로 발생하였다. 따라서 산불 발생 확률은 낮지만 실제로 산불은 높은 빈도로 발생하였으므로 확률 모형의 분류 성능이 좋지 않다는 것을 의미하게 되고 $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$의 ROC curve의 auc 값이 작게 나온다.

$P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$의 앞의 15번, 2번 지역뿐만 아니라 나머지 지역 모두에 대한 확률 좌표 그래프를 보면 대부분의 y값(확률값)이 높은 좌표의 x값에 세로선이 겹쳐져서 그려져 있는 것을 볼 수 있다. 이는 $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$의 성능이 20개 지역 전체에 대해 우수하다는 결과를 그래프로 보여주는 것이다. 20개 지역에 대한 $P_{dl,mean}(Z_t=1|\boldsymbol{X_t})$의 확률 좌표 그래프는 <부록 2>의 <그림 25>~<그림 44>에 있다.

또한 확률 좌표 그래프를 보면 좌표들이 일정한 형태를 가지고 그려지는 것을 볼 수 있다. 형태를 살펴보면 비슷한 간격을 사이에 두고 좌표의 y값이 증가했다가 다시 감소하는 형태가 그래프에 반복적으로 나타나고 있다. 이 때 x값이 0에서 1000 사이인 구간에서 3번 정도 좌표의 y값이 증가했다가 감소하는 형태가 반복되므로 약 x값이 330, 즉 약 1년 정도가 산불 발생 확률의 증감형태 사이의 간격이라고 할 수 있다. 이는 1년 중 산불 조심기간인 봄(2, 3, 4, 5월)에 산불 발생 확률이 크게 증가하는 현상이 좌표의 y값(확률값)의 변화로 나타난 것으로 보인다.

2. 산불 피해규모 예측 모형의 결과와 고찰

본 연구의 t시점에서 산불이 발생하였을 때 기대되는 피해규모 예측 모형으로 가장 일치도가 높고 성능이 우수한 $E_{dl,mean}(Y_t|Z_t=1,\boldsymbol{X_t})$을 최종적으로 선택하였다. 반응변수가 7개의 단계로 이루어져 있으므로 단순한 분류의 일치 확률인 1/7(≒0.142857)의 2배보다 모형의 예측값과 실제값의 일치도가 크게 나왔지만, 다음과 같은 개선을 통해 더 성능이 우수한 모형을 적합하는 것이 가능할 것이다.

본 연구에서 사용한 전국 산불발생 통계자료의 산불 피해면적 정보는 산불피해면적의 전체 크기이다. 이를 기상 데이터를 이용해 예측하려고 하면, 확산의 속도, 방향, 크기 등이 산불의 전체 크기에 큰 영향을 미치는데 이를 결정하는 산불 발생지의 경사와 같은 지형, 주변 나무의 종류 및 밀도 등의 숲의 구성과 같은 요인이 고

려되지 않는다. 따라서 기상 데이터와 지형, 숲의 구성 등 산불의 확산과 그 총 피해면적에 영향을 미치는 모든 요인을 설명변수로 이용하여 피해면적 예측 모형을 적합한다면 더 성능이 좋은 모형이 될 것이다.

3. 기상 데이터가 주어졌을 때 기대되는 산불 피해규모 예측 모형의 결과와 고찰

최종적으로 선택한 산불 발생 확률 모형은 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$이고 산불이 발생했을 때 기대되는 피해규모 예측 모형은 $E_{dl,mean}(Y_t|Z_t = 1, \boldsymbol{X_t})$이다. 따라서 본 연구에서 제안하고자하는 특정일(t시점) 이전 14일의 기상 데이터($\boldsymbol{X_t}$)가 주어졌을 때 기대되는 산불 피해규모 예측 모형은 $E_f(Y_t|\boldsymbol{X_t})$이라 나타내고 다음의 식 (9)와 같다.

$$E_f(Y_t|\boldsymbol{X_t}) = E_{dl,mean}(Y_t|Z_t = 1, \boldsymbol{X_t})P_{dl,mean}(Z_t = 1|\boldsymbol{X_t}) \quad (9)$$

V. 결론

본 연구는 산불 위험관리의 측면에서 한계점이 있는 산불이 발생했을 때 기대되는 피해면적이 아닌 이전의 14일 동안의 기상 데이터가 주어졌을 때 특정한 하루의 산불 피해규모의 기댓값을 예측하는 모형을 제시하였다. 그 결과 많은 지역에서 이 모형을 이용해 상시적으로 특정지역의 산불피해 예측규모를 산정할 수 있었고, 이는 새로운 산불위험 지수로 사용할 수 있을 것이다.

산불 발생 확률 예측 모형과 산불이 발생하였을 때 기대되는 산불 피해규모 예측 모형 모두에서 평균값을 선택하는 방식으로 기상 관측 자료를 가공하여 이용하였을 때 모형의 성능이 가장 우수한 것으로 나타났다. 그리고 20개 지역에 대해 산불 발생 확률 예측 모형의 적합에 있어 정규화 로지스틱 회귀분석 모형, 단층 신경망 모형 및 다층 신경망 모형의 성능을 비교하였고 많은 지역에서 다층 신경망 모형을 적합하기 위해 사용한 딥 러닝(Deep-learning) 알고리즘을 활용하였을 때 그 성능이 가장 우수했다. 산불이 발생하였을 때 산불의 피해규모 예측 모형의 적합에서는 k-최근접 이웃 알고리즘 및 다층 신경망 모형의 성능을 비교하였고 그 중 다층 신경망 모형을 적합하기 위해 사용한 딥 러닝 알고리즘을 활용하였을 때 성능이 가장 우수했다. 따라서 두 모형의 적합에 있어 딥 러닝 알고리즘이 설명변수의 시간적, 공간적 비선형관계와 정보를 효과적으로 축약하여 성능이 좋은 다층 신경망 모형을 적합하였다고 할 수 있다.

하지만 딥 러닝 알고리즘을 이용하여 모형을 적합하면 모형의 해석이 거의 불가능하여 산불을 관리하는 실무자에게 예측 확률과 피해규모 외의 다른 정보를 주지 못한다는 단점이 있다. 이러한 점에서 성능은 조금 떨어지지만 해석이 보다 쉬운 정규화 로지스틱 회귀분석 모형이 유용하게 사용될 수 있을 것이다.

산불이 발생하였을 때 산불의 피해규모 예측 모형의 적합에서 회귀분석을 하지 못하고 대신 분류(Classification) 분석 방법을 이용하였다. 그 이유는 반응변수가 매우 꼬리가 두터운 분포를 가지므로(<그림 2> 참고) 선형모형이나 비선형 모형을 이용한 모형의 적합이 어렵기 때문이다. 따라서 평균이 존재하지 않는 반응변수에 대한 회귀분석 방법을 고려해야 할 것이다. 그 중 하나의 후보 모형으로는 분위수 추정 모형을 생각해 볼 수 있다.

전국 산불발생 통계자료에 기록된 정보 중 하나인 산불의 발생 원인을 고려하여 설명변수로 포함한 연구가 진행되지 못했다. 연구에서 사용된 기상 데이터는 자연적 요인이기 때문에 산불의 발생과 그 크기에 영향을 줄 수 있는 인위적 요인을 같이 고려하여 모형을 제작하면 더 좋은 성능의 산불 발생 확률 예측 모형이나 산불 피해면적 예측 모형의 적합이 가능하다고 생각된다. 또한 산불 피해면적 예측 모형을 연구할 때 경사, 숲의 구성과 같은 지형적 요인들이 고려되어 설명변수로 포함되지 않았다. 지금까지 개발된 산불 위험 지수 모형은 기상조건과 함께 고도, 방위와 같은 지형조건 및 임상조건들을 모두 고려했고, 그것이 산불 위험 지수의 예측력을 높여주는데 성공적이었다고 알려져 있다. 따라서 모형의 적합에서 기상 데이터와 함께 지형조건과 임상조건들을 설명변수로 포함하여 모형을 적합한다면 성능이 좋은 모형을 기대할 수 있을 것이다.

<참고문헌>

김승수·이종호·이명욱(2013), “50년간 통계분석을 통한 국내 산불과 기상과의 관계,” 『한국방재학회논문집』, 13(5): 225~231, 한국방재학회.

박종길·정우식·김은별·장현준(2014), "산불발생에 영향을 미치는 기상현상의 분석에 관한 연구," 『한국방재학회 학술발표대회논문집』2014: 376, 한국방재학회.

박홍석·이시영·채희문·이우균(2009), “캐나다 산불 기상지수를 이용한 산불발생확률모형 적합 : 강원도 지역 산불발생을 중심으로,” 『한국방재학회논문집』, 9(3): 95~100, 한국방재학회

이병두·정주상(2006), “1970-2005년 동안의 산불 발생건수 및 연소면적에 대한 시계열모형 추정,” 『한국임학회지』, 95(6): 643~648, 한국임학회.

이시영·한상열·원명수·안상현·이명보(2004), “기상특성을 이용한 전국 산불발생확률모형 적합,” 『한국농림기상학회지』, 6(4): 242~249, 한국농림기상학회.

최관·한상열(1996), “기상자료를 이용한 산불발생확률모형의 적합,” 『한국임학회지』, 85(1): 15~23, 한국임학회.

Cortez P. and Morais A.(2007), “A Data Mining Approach to Predict Forest Fires using Meteorological Data,” *New Trends in Artificial Intelligence*, 13: 512~523, Portuguese Conference on Artificial Intelligence, Guimaraes, Portugal.

http://www.kma.go.kr/HELP/basic/help_01_01.jsp (2016. 7. 14.).

Larjavaara, M., Kuuluvainen, T., Tanskanen, H., and Venalainen, A.(2004), “Variation in forest fire ignition probability in Finland,” *Silva Fennica*, 38(3): 253~266.

Safi, Y., and Bouroumi, A.(2013), "Prediction of forest fires using artificial neural networks," *Applied Mathematical Sciences*, 7(6): 271~286.

Van Wagner, C. E., & Forest, P.(1987), "Development and Structure of the Canadian Forest Fire Weather Index System," *Forestry Technical Report*, 35: 35, Canadian Forest Service.

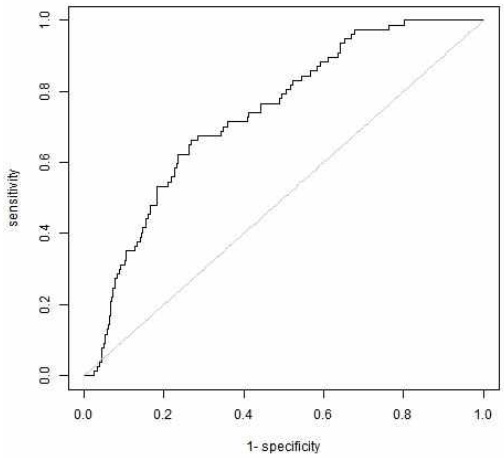
Xiao, F.(2012), "Forest Fire Disaster Area Prediction Based on Genetic Algorithm and Support Vector Machine," *Advanced Materials Research,* 446~449(4): 3037~3041, TRANS TECH PUBLICATION LTD.

Xie, D.W., and Shi, S.L.(2014), "Prediction for Burned Area of Forest Fires Based on SVM Model," *Applied Mechanics and Materials,* 513~517(5): 4084~4089, TRANS TECH PUBLICATIONS LTD.

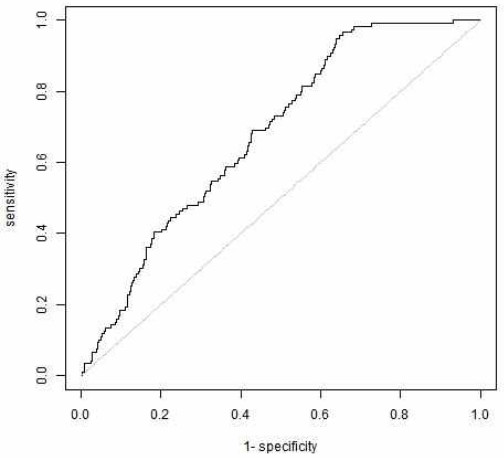
<부록 1>

<그림 5>~<그림 24>: 20개 지역의 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$의 ROC curve

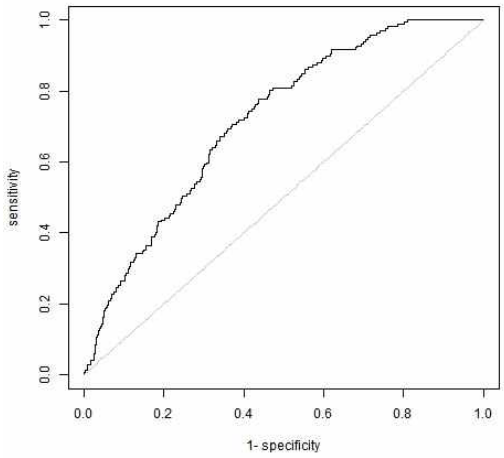
<그림 5> 1번 지역



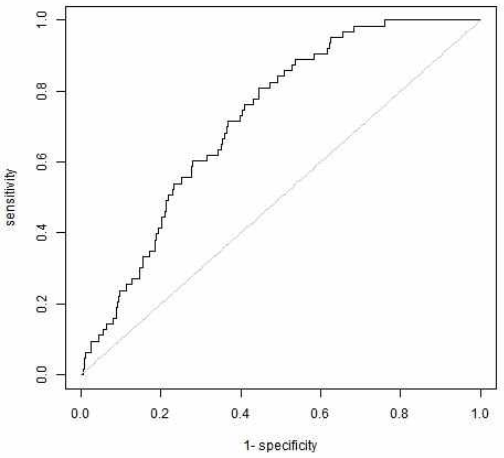
<그림 6> 2번 지역



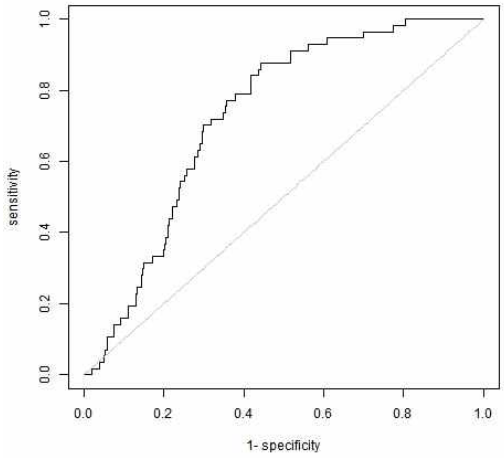
<그림 7> 3번 지역



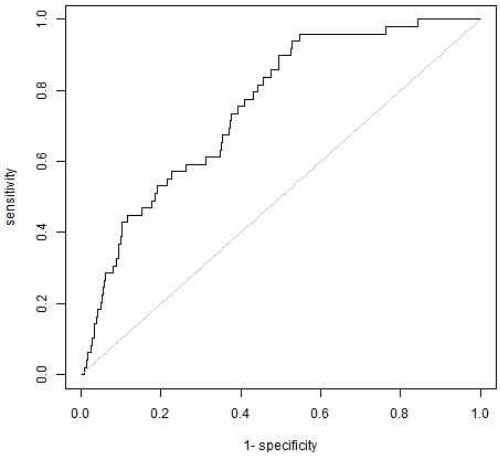
<그림 8> 4번 지역



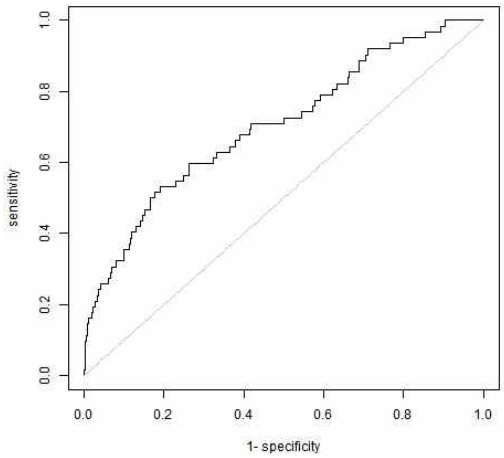
<그림 9> 5번 지역



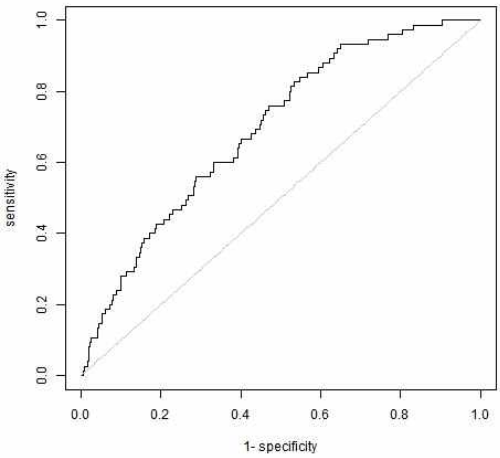
<그림> 10 6번 지역



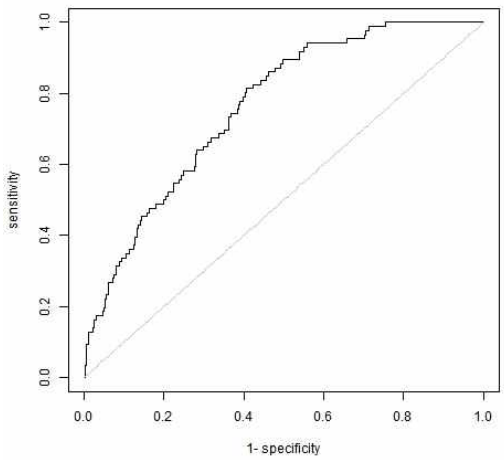
<그림 11> 7번 지역



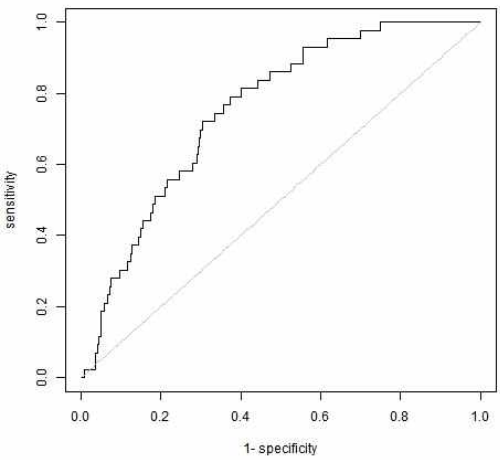
<그림 12> 8번 지역



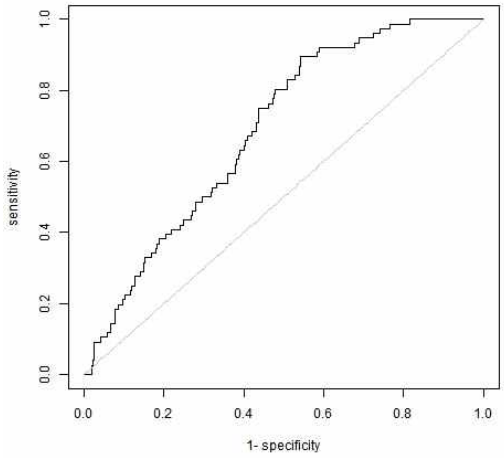
<그림 13> 9번 지역



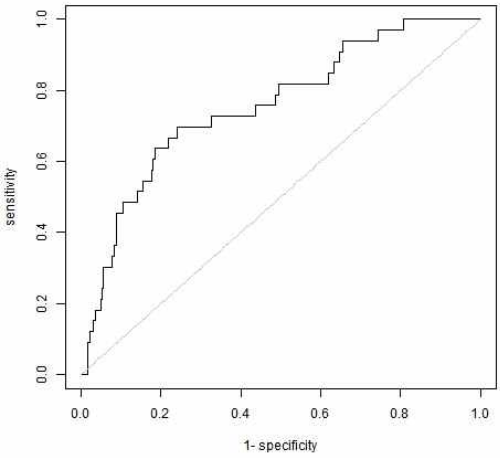
<그림 14> 10번 지역



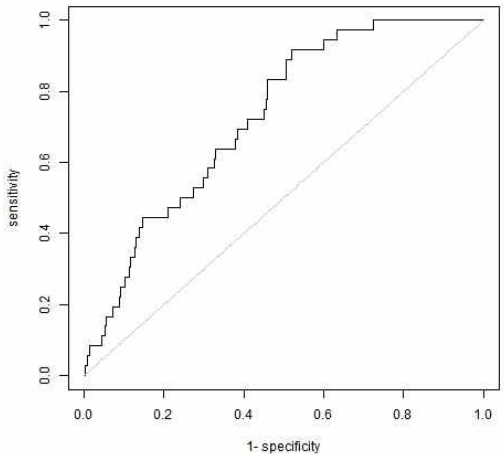
<그림 15> 11번 지역



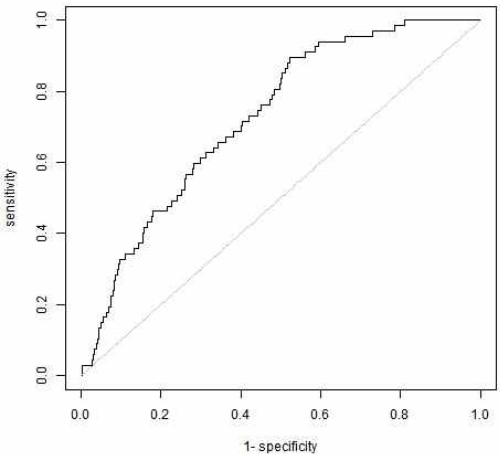
<그림 16> 12번 지역



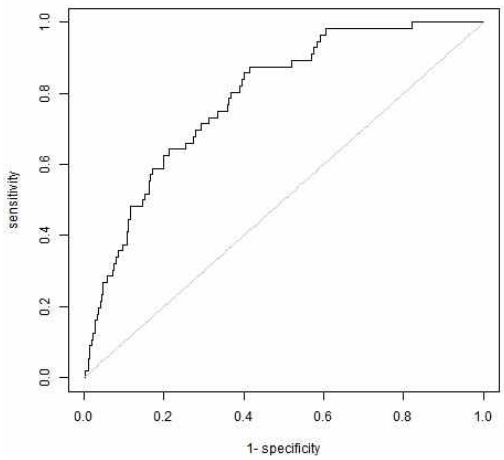
<그림 17> 13번 지역



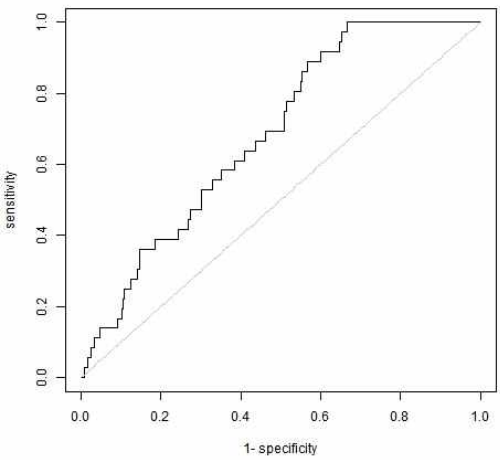
<그림 18> 14번 지역



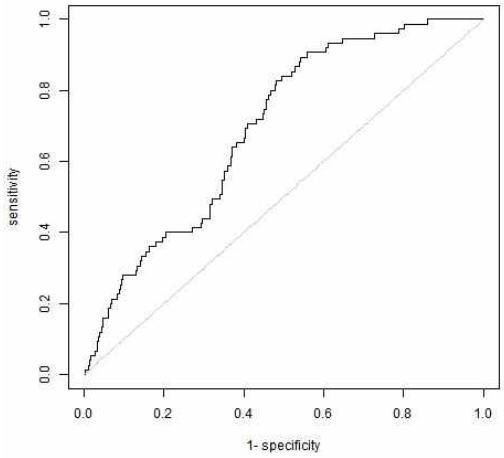
<그림 19> 15번 지역



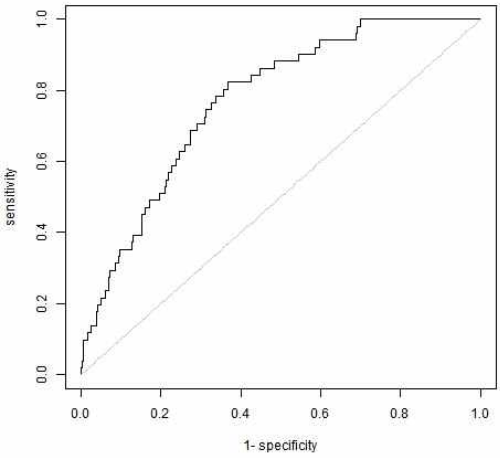
<그림 20> 16번 지역



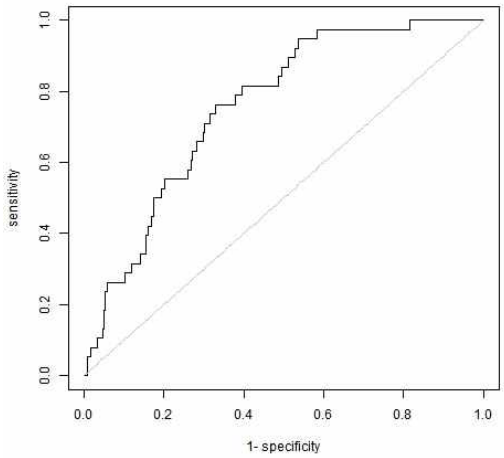
<그림 21> 17번 지역



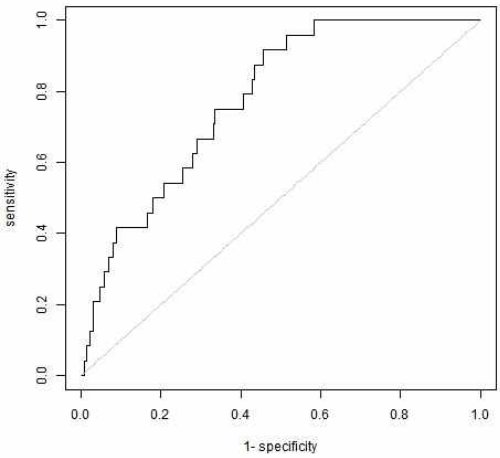
<그림 22> 18번 지역



<그림 23> 19번 지역



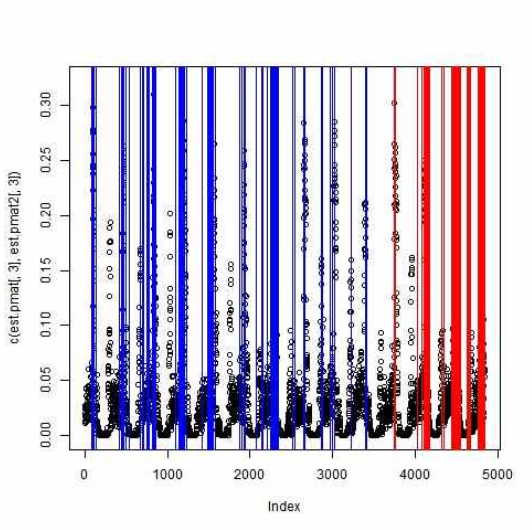
<그림 24> 20번 지역



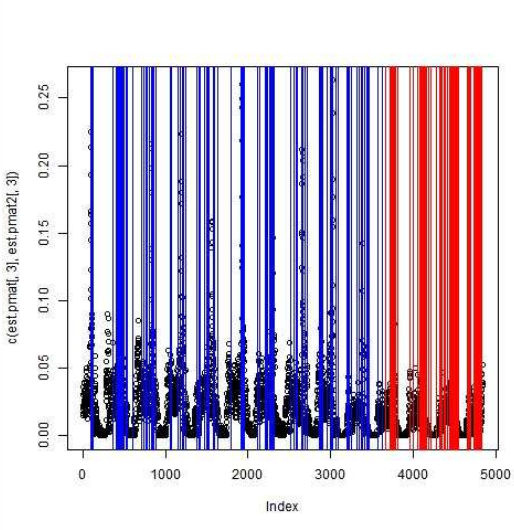
<부록 2>

<그림 25>~<그림 44>: 20개 지역에 대한 $P_{dl,mean}(Z_t = 1|\boldsymbol{X_t})$의 확률 좌표 그래프

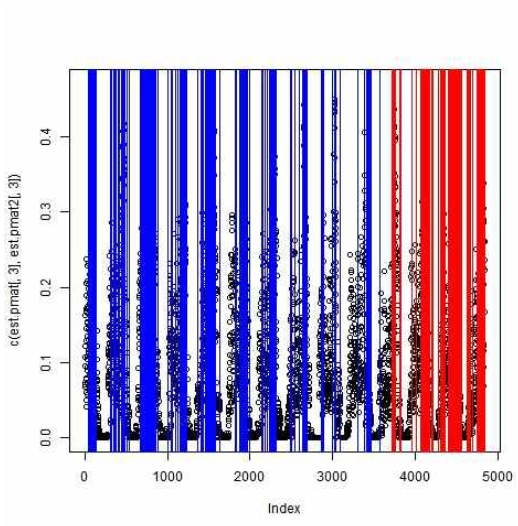
<그림 25> 1번 지역



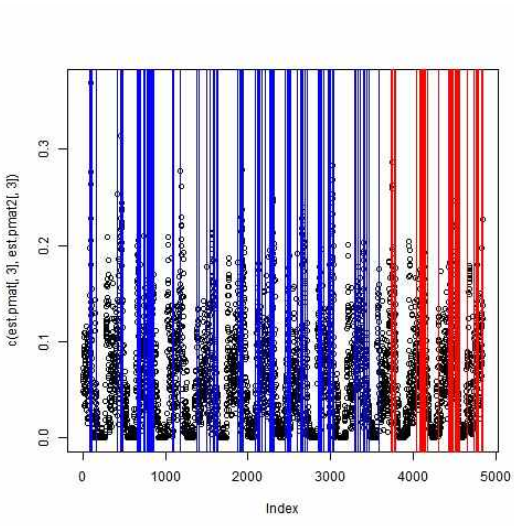
<그림 26> 2번 지역



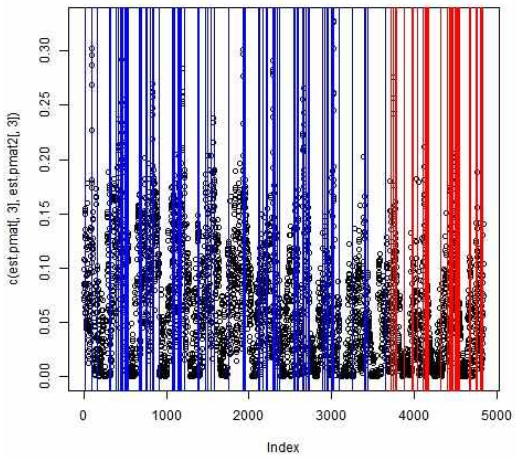
<그림 27> 3번 지역



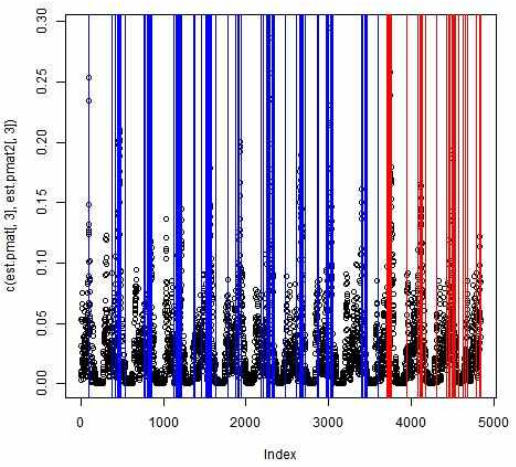
<그림 28> 4번 지역



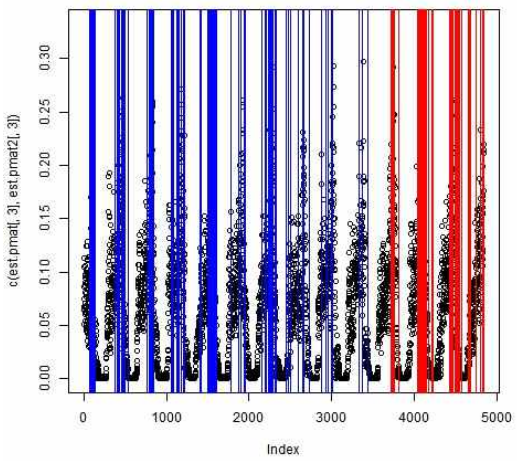
<그림 29> 5번 지역



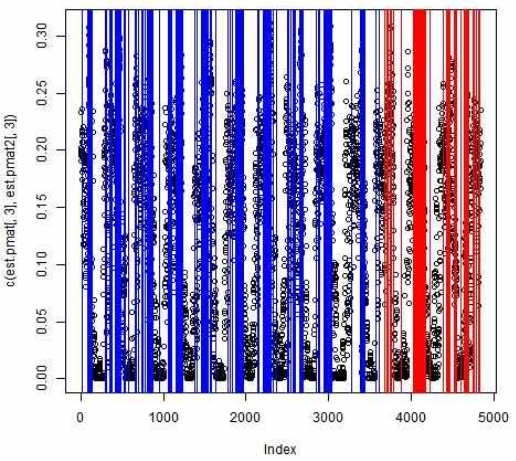
<그림 30> 6번 지역



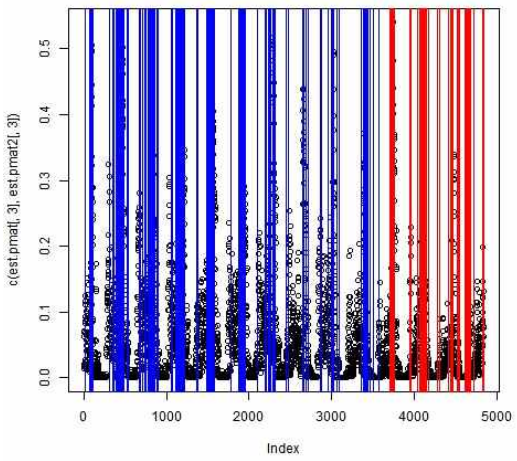
<그림 31> 7번 지역



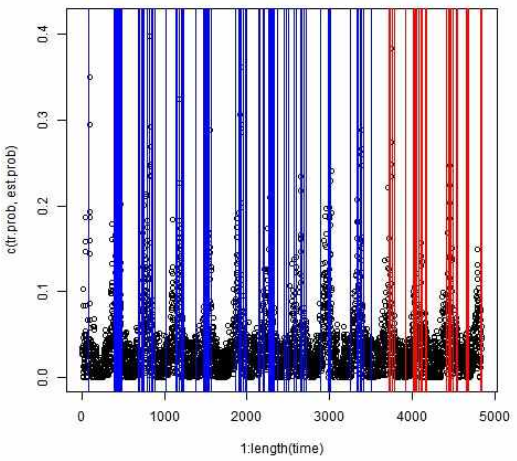
<그림 32> 8번 지역



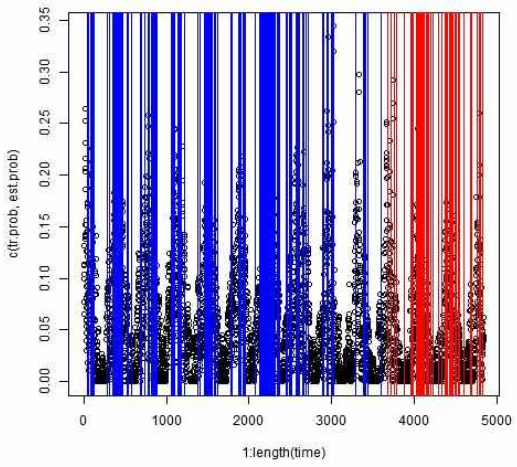
<그림 33> 9번 지역



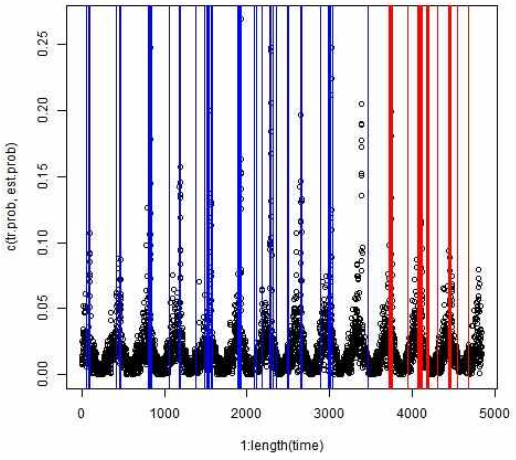
<그림 34> 10번 지역



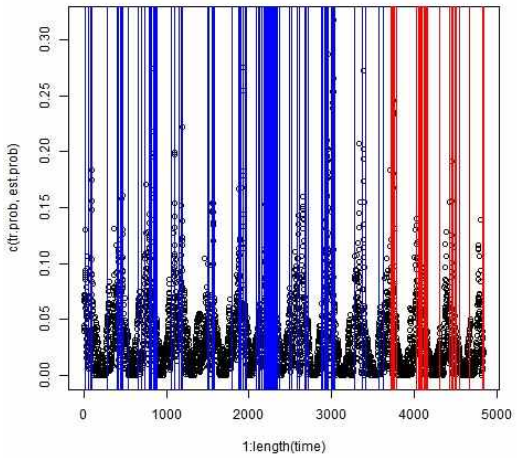
<그림 35> 11번 지역



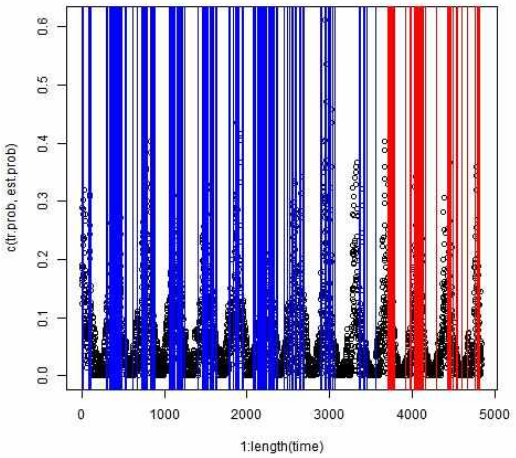
<그림 36> 12번 지역



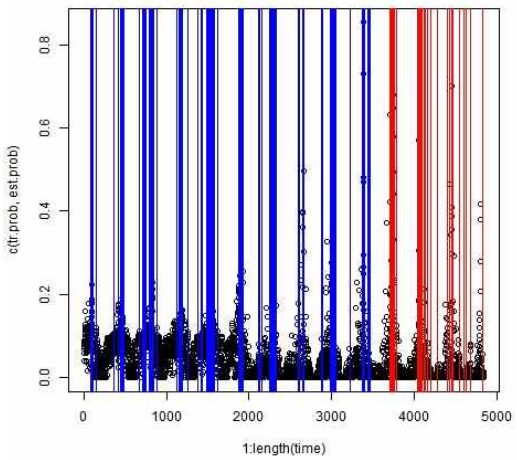
<그림 37> 13번 지역



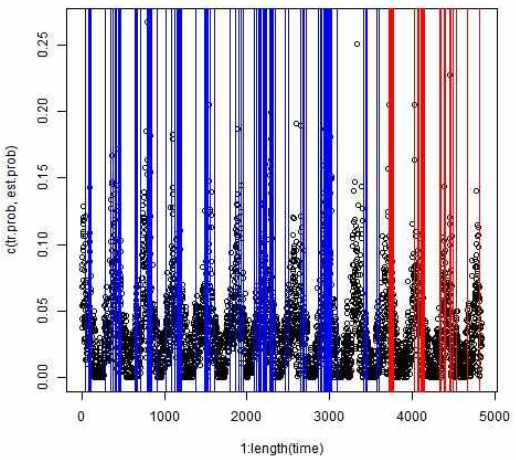
<그림 38> 14번 지역



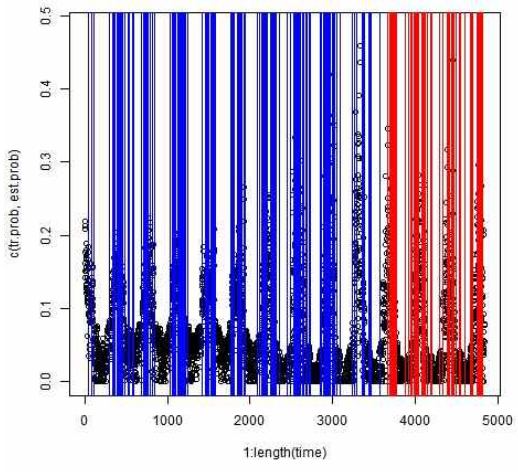
<그림 39> 15번 지역



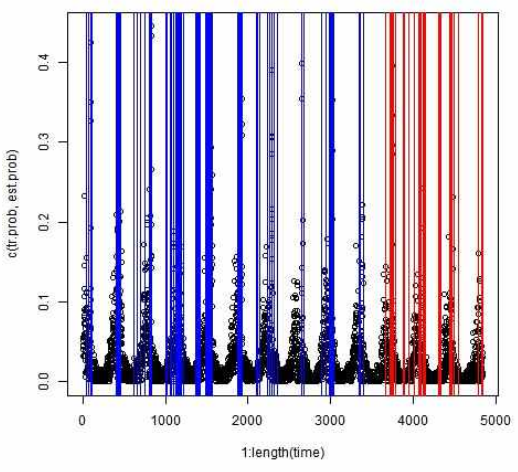
<그림 40> 16번 지역



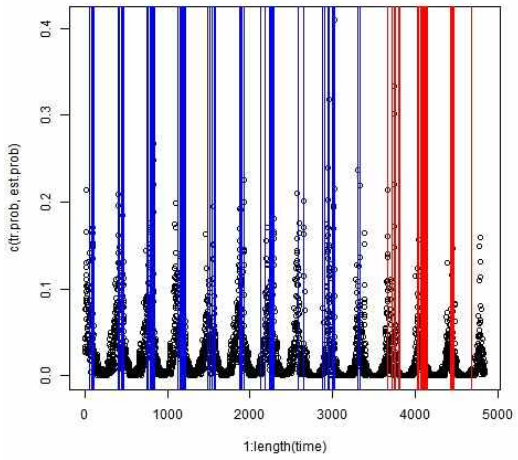
<그림 41> 17번 지역



<그림 42> 18번 지역



<그림 43> 19번 지역



<그림 44> 20번 지역