creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

농학석사 학위논문

가시광선 및 근적외선 분광분석법을 이용한 붉가시나무 종자의 충해 정확도 검정

Accuracy Test on Insect-damaged seeds of *Quercus acuta* using Visible-Near Infrared Spectroscopy

2023년 02월

서울대학교 대학원

농림생물자원학부 산림환경학 전공

조계홍

가시광선 및 근적외선 분광분석법을 이용한 붉가시나무 종자의 충해 정확도 검정

Accuracy Test on Insect-damaged seeds of *Quercus acuta* using Visible-Near Infrared Spectroscopy

지도 교수 강규석

이 논문을 농학석사 학위논문으로 제출함
2022년 12월

서울대학교 대학원
농림생물자원학부 산림환경학 전공
조계홍

조계홍의 농학석사 학위논문을 인준함
2023년 01월

위 원 장 김현석 (인)

부위원장 강규석 (인)

위 원 박필선 (인)

초 록

국내외적인 탄소중립 흐름에 따라 탄소흡수원의 확충을 위한 기후변화 대응 활엽수종의 필요성이 높아지고 있다. 붉가시나무는 기후변화에 의해 한반도 해안 지역을 중심으로 내륙까지 잠재 생육지역이 확장될 것으로 기대되는 난대상록 활엽수종이다. 차후 조림수요를 위해 효율적으로 종자를 생산하기 위해서는 저비용, 비파괴적인 분광분석법을 연구할 필요가 있다. 그러나 국내의 경우 근적외선을 이용한 산림종자의 연구는 아직 미비하고, 해외 또한 난대상록 활엽수종의 종자에 대한 품질 분류 연구는 미비한 상황이다. 따라서 밤 및 석류 등을 대상으로 한 사전연구를 바탕으로, 근적외선 분광분석법을 이용하여 400~1000nm 구간대에서 붉가시나무 종자의 충해 여부를 판단하고자 본 연구를 실시하였다. 제주도 서귀포시에 조성된 붉가시나무 채종원에서 550립의 종자를 채집하였다. 채집한 종자의 충해 여부를 시각적인 방법으로 건전종자 362립, 충해종자 178립으로 구분하였다. 550립의 종자들은 각각 calibration을 위한 366립(건전 246/충해 120)과 validation을 위한 184립(건전 116/충해 58)으로 무작위로 나누어졌다. Ocean-optics사의 Flame-S-Vis-Nir 분광분석기와 적분구(ISP-REF)로 종자의 각 방위를 10번씩 측정한 값을 평균하여 총 550개의 근적외선 파장데이터를 획득하였다. 분광분석 데이터의 오류를 줄이기 위하여, 1차 미분, MSC, 표준정규화, Savitsky-Golay 필터 등의 데이터 전처리를 적용하고, PLS-DA, OPLS-DA 다변량 분석법을 적용하였다. 정확도 및 정밀도를 비롯한 모델 평가 지표 및 Variable Importance in Projection을 이용하여 파장 사용 구간, 전처리 및 다변량 분석법에 따른 모델성능을 비교하였다. 근적외선 파장측정 결과, 붉가시나무 종자의 근적외선 파장은 660-720nm에서 강한 흡광지점을 보여주었고, 900~1000nm에 걸쳐 미약하게 아래로 향하는 흡광

지점을 보여주었다. 다변량 분석 및 모델 평가 결과, 가장 우수하게 충해여부를 분간하는 모델은 400~1000nm 구간대에 SG처리를 하여 PLS-DA를 적용한 모델로, 정확도가 0.8641, 정밀도 0.8603, F1 Score 0.9035, 그리고 G mean 0.8093($p<0.05$)이었다. 모델의 성능에는 660-720nm와 960-1000nm의 흡광지점이 유의미한 영향을 주었다. 후자는 충해 여부가 수분과 탄수화물에 영향을 준 것으로 추정되었다. 전자의 경우 엽록소, 색상변이 등으로 인한 영향이 낮은 분별력에 영향을 주었으나 근적외선 파장대와 같이 쓰였을 때 모델의 성능에 영향을 미쳤다. 흡광지점에 대한 본 연구결과는 추후 더 넓은 파장대나 다차원 분광분석을 이용한 후속연구의 조사 범위를 좁히는 데 활용될 수 있을 것이다. 본 연구의 정확도를 개선하기 위해서는 x-ray 등 다른 형태의 자료를 병용하거나 기존 충해종자의 분류군을 세분화하거나 데이터 불균형성을 해소할 필요가 있다고 판단하였다.

주요어 : 근적외선 분광분석, 붉가시나무, 다변량분석, 충해, 모델평가
학　번 : 2021-29417

목차

초 록 i
표 목차 iii
그림 목차 iv
제 1 장 서 론 5
제 1 절 연구의 배경 및 필요성 5
제 2 절 연구 목적 7
제 3 절 용어 정의 7
제 2 장 연 구 사 11
제 1 절 근적외선 품질분류 11
제 2 절 붉가시나무 12
제 3 장 연구 방법 14
제 1 절 연구재료 14
제 2 절 충해종자 분류 16
제 3 절 종자 전처리 17
제 4 절 NIR 파장 측정 18
제 5 절 데이터 분석 21
제 1 항 데이터 전처리 21
제 2 항 다변량 분석기법 적용 23
제 3 항 모델 성능 평가지표 25
제 4 장 연 구 결 과 27
제 1 절 NIR 파장의 형태 분석 27
제 2 절 충해종자 검정 모델 간 비교 28
제 5 장 고 찰 44
제 1 절 주성분분석 상에서 이상치 처리 44
제 2 절 파장변수대에 대한 고려 45

제 3 절 전처리 및 다변량 분석기법의 처리 효과 47
제 4 절 연구의 의의 및 한계 49
제 5 절 후속 연구에 대한 제언 50
제 6 장 결 론 52
Abstract 54
참고문헌 56

표 목차

Table 1. Number of acorns collected from each family 15
Table 2. Evaluation of accuracy and precision..... 26
Table 3. Evaluation index for raw spectra 30
Table 4. Evaluation index for standard normal variate process ed spectra 33
Table 5. Evaluation index for multiplicative scatter correction processed spectra 35
Table 6. Evaluation index for 1st derivative processed spectra 37
Table 7. Evaluation index for Savitsky−Golay filter processed spectra 39

그림 목차

Figure 1. Location of the seedling seed orchard of *Quercus acuta* in Jeju 15
Figure 2. Picture of the caterpillar of *Curculio dentipes* 16
Figure 3. Water bath treatment for temperature management 17
Figure 4. Acorn spectroscopy using integrating sphere..... 20
Figure 5. Whole VIS-NIR raw spectra from 400 to 1000nm for acorns of *Quercus acuta* 27
Figure 6. Average of wavelength between sound and damage acorns from 400 to 1000nm 28
Figure 7. Principal component analysis for sound and damaged acorns in wave range from 780 to 1000nm 29
Figure 8. Spectra processed with standard normal variate 400~1000nm 31
Figure 9. Spectra processed with multiplicative scatter correction at 780~1000nm 34
Figure 10. Spectra processed with 1st derivative 36
Figure 11. Spectra processed with Savitsky-Golay filter 38
Figure 12. Top 5 evaluated combinations of analyzation methods 40
Figure 13. 400~1000nm Savitsky-Golay filter (SG) processed model's PC score plot 41
Figure 14. Y-predicted plot 42
Figure 15. Variable Important for the Projection..... 43
Figure 16. Regression Coefficient 43
Figure 17. Schematic of outlier detection 44
Figure 18. PLS weight of 400~1000nm Savitsky-Golay filtered PLS-DA(SGWP) model's 1st PC 46
Figure 19. Specificity by preprocessing methods 48

제 1 장 서 론

제 1 절 연구의 배경 및 필요성

전세계적으로 탄소중립 흐름이 이어지면서 산림의 조성 및 관리가 매우 중요해졌다. 우리나라에서도 2050 산림부문 탄소중립 추진전략안을 통해 각 분야에서 탄소중립 추진 및 저탄소정책(안)을 제시하였다(미래전략과, 2020). 이는 경제구조의 저탄소화, 신유망 저탄소 산업 생태계 조성, 탄소중립 사회로의 공정전화, 탄소중립 제도적 기반 강화 등 경제 및 사회정책과 관련된 크게 4개의 대분야로 구성되어 있다. 여기에 장기적인 저탄소발전을 내다보는 장기저탄소발전전략은 크게 5가지로 구성되어 있는데, 이중 유휴토지 신규조림 및 산림연령 저하 등을 통한 산림, 갯벌, 습지 등의 탄소 흡수능 향상이 들어있다.

탄소흡수능 향상을 위한 방안으로 주로 탄소흡수원이 언급되며, 탄소흡수원의 확충은 노령림 교체 및 가시나무 등으로의 수종갱신 및 간벌 등을 통해 이뤄질 예정이다(미래전략과, 2020). 이러한 탄소중립 흐름에 맞춰 유휴토지 등에는 2050년까지 약 1500ha, 도시림 및 생활숲 및 기타 등등까지 포함하면 약 4.4만ha에 달하는 탄소흡수원이 조성될 예정이며, 이에 따라 기후변화 대응 활엽수종의 조림 수요 또한 늘어날 것이다(미래전략과, 2020).

가시나무류는 기후변화 대응 활엽수종에 부합하는 탄소흡수원 수종이다. 예를 들어 붉가시나무(*Quercus acuta* Thunb.)의 경우, 현재는 난대성 상록 활엽수종으로서 한반도 남부지역의 제주도, 완도 등의 전라남도 남부 지역에 자생하고 있으며 해발고 170-500m 대의 산록, 계곡의 양지에서 주로 자생하고 있다(이정현 등, 2010).

기후변화가 진행됨에 따라 미래 잠재생육지역이 한반도 남해, 서해, 동해안 지역을 중심으로 내륙으로 확장될 것으로 기대되고 있다(신만석 등,

2018). 이러한 미래의 조림수요에 따른 종자 공급을 위하여 제2차기 채종원 조성 및 관리 추진계획에 따라 현 19ha면적의 가시나무류 채종원(참가시나무, 종가시나무, 붉가시나무, 개가시나무)에 추가하여 28ha를 추가로 신규 조성하여 기존 채종원 면적의 2.5배로 늘릴 계획이다(국립산림품종관리센터, 2021).

종자들의 대량 공급에 있어 종자의 품질은 효율에 직접적으로 연관이 되어 있기 때문에 중요시된다(Novikov and Novikova, 2018). 왜냐하면 종자의 품질은 종묘상 구성 비용, 묘목의 품질, 조림의 성공 여부 등에 관련되어 있으며 기후변화 등으로 인해 다변화하는 환경조건 하에서도 우량한 품질, 빠른 식물의 생장에 긍정적인 영향을 주기 때문이다(FAO, 2010). 종자의 품질을 구성하는 요소 중에는 발아율, 활력, 충해 여부 등이 존재한다(FAO, 2010).

붉가시나무의 경우 도토리의 높은 충해 비율이 문제가 되고 있다. 격년 결실 주기가 뚜렷한 붉가시나무 특성 상(Hirayama et al., 2012), 종자가 덜 열리는 해에 충해 피해를 더 받을 수밖에 없다. 한반도 남부의 붉가시나무는 9~10월 사이의 결실종자 중 절반 이상이 충해 피해를 받고 있어(Kim and Park, 2021), 건전종자와 충해종자의 온전한 구분의 필요성이 높아지고 있다.

도토리의 충해를 분류하는 기존의 방법으로는 종자를 물에 띄워서 밀도가 낮아서 물에 뜨는 종자들을 제거하는 수선법, 즉 floatation method를 사용해왔다(Gribko and Jones, 1995). 수선법만으로는 충해 피해를 입은 종자를 전부 걸러낼 수 없어 결국 종자 보관단계에서 벌레가 발아하는 문제가 발생하기 때문에(Connor, 2004).

국립산림품종관리센터 내에서는 종자 처리 및 정선, 종자품질검정 등 자동화된 스마트 종자처리 시설을 설치하고 있다(산림청, 2021). 하지만 종자검정 및 정선에 있어 기존의 방법, 즉 테트라졸륨, 풍선법, 절단검사 등은 종

자를 파괴하여 품질 검사 이후 종자를 쓸 수 없는 문제가 있으며(Novikov, 2019), 기계화를 이용한 대량생산 공정에는 이용되기 어려웠다(Novikov et al., 2020). 그리하여 근적외선 분광분석법을 이용한 종자검사 방법은 저비용, 비파괴검사로 각광받아왔으며(Novikov, 2019), 동시에 여러 형질들을 검사할 수 있는 방법으로 쓰이기도 하였다(Shenk and Westerhaus, 1993).

제 2 절 연구 목적

본 연구는 종자의 품질에 영향을 미치지 않으면서도 종자 검사가 가능한 근적외선 분광분석 방법을 이용하여 붉가시나무의 충해 피해를 검정할 수 있는지를 구명하고 이를 바탕으로 분별 모델을 만들고자 하였다. 이를 위하여 우선적으로 접근성이 가장 높은 400~1000nm 구간의 가시광선 영역과 근적외선 일부를 바탕으로 파장, 종자 품질 분류 데이터를 이용하여 종자 품질의 검정 및 유의미한 분류 모델 확보가 가능한지를 검증하였다. 그리고 모델간 비교를 하여 정확도 및 모델 평가 지수들이 높은 데이터 전처리, 다변량 분석기법 조합에 따른 구분 정도를 구명하였다. 마지막으로는 종자의 분류가 유의미하게 이뤄지는 가시광선 및 근적외선 파장대의 파장변수들을 추출하였다.

제 3 절 용어 정의

데이터 분석은 통계적으로 오류를 수정하는 데이터 전처리, 전처리된 데이터와 분류 데이터를 이용하여 분류 모델을 만드는 다변량 분석기법 단계로 구성되어 있다. 근적외선 분광분석에서 데이터 전처리는 분광분석 시 발생하는 오류를 통계적으로 제거하는 과정을 말한다. 근적외선 분광분석 시 주로 발생하는 오류에는 크게 두가지가 있다.

우선, 기준값(Baseline shifts)라 불리는 오류가 있다. 다른 오류로는 non-

linearity라 불리는 오류이다. 이 두가지 오류는 재료 표면의 상태나 미세입자, 미생물, 빛의 산란 경로 등 수많은 요소에 의해 발생한다. 특히 생물에 관련된 재료의 경우, 무수히 많은 요소에 의해 오류가 발생하기 때문에 전처리를 이용하여 이러한 오류를 없애 줘야 한다. 또한 이러한 전처리 방법들을 이용하여 흡광지점(peak)을 분리해내고 강조를 하는 효과를 가질 수 있다(Rinnan et al., 2009).

PCA, PLS-DA, OPLS-DA는 데이터 분석에 있어 차원을 축소하여 이를 기반으로 데이터를 분석하는 기법이다. PCA는 Principal Component Analysis의 준말로, 매우 많은 변수를 벡터를 이용하여 차원을 축소하여 몇 개의 주성분으로 만드는 방법이다. 이 방법을 이용함으로서, 변수들 간에 존재하는 공분산성을 제거하고 변수의 개수를 축소함으로서 다변량 분석이 용이하게끔 해준다.

PLS-DA는 Partial Least Square Discriminant Analysis의 준말이다. Partial least square란 부분 최소 제곱법을 의미한다. 무수히 많은 변수들이 존재하는 데이터에 대하여 X값과 Y값 간의 공분산을 고려하여 중요한 변수들을 추출하여 몇 개의 pls변수들로 차원을 축소시키는 방법이다. 데이터의 분산을 설명하는데 있어 X값들 간의 공분산, X와 Y값 간의 공분산을 고려하기 때문에, 데이터를 설명하는 데 있어 예측변수와 설명변수 간의 상관관계를 포함시킬 수 있다. 그 결과, 이러한 상관관계를 포함한 주성분들이 만들어지며, 최종적으로 Y값은 해당 주성분들로 구성된 regression식에 의해 설명이 된다.

Discriminant analysis는 두 개 혹은 그 이상의 범주형 데이터로 구성된 Y값에 따라 데이터를 분류하는 방법으로 -1,1 혹은 0,1로 구성된 Y 가변수를 사용하였다. 만약 0,1로 구성된 Y 가변수를 사용하였을 경우 Y값들은 0 그리고 1근접한 값들을 가지게 될 것이다. 보통 threshold, 즉 기준을 설정하여 기준에 따라 데이터성분들의 예상된 범주를 결정하게 되는데, 이 경우 기

준값은 Y=0.5가 된다. 즉, PLS를 통해 만들어진 회귀식에 의해 나온 예상값이 0.5를 기준으로 그 이상이면 1, 그 이하면 0으로 분류된다고 볼 수 있다.

OPLS-DA는 Orthogonal Partial Least Square Discriminant Analysis의 준말이다. PLS-DA의 방법의 일종으로 PLS-DA 분석 시 X, Y 뿐만 아니라 수직성분을 집어넣은 분석방법이다. 이 경우, PLS에서 본래 X-Y간의 공분산이 반영된 주성분들로 설명하던 부분을 X-predictive와 X-orthogonal이라는 두가지 형태의 주성분으로 나누어서 설명하게 된다. X-predictive의 경우, X-Y간의 공분산의 최대치가 한 번에 반영된 주성분으로, 보통 사용하는 Y-가변수가 1개일 경우 마찬가지로 1개로 고정된다. X-orthogonal의 경우 이러한 X-Y간의 공분산이 제외되어 X변수들간의 분산만을 설명하는 변수들이다(Trygg and Wold, 2002).

정확도(accuracy)는 모델에 의한 분류가 실제 분류와 일치하는 비율을 나타내는 것으로, (TP+ TN)/(TP+ TN+ FP+ FN)로 계산이 가능하다. 재현율(recall)은 실제 positive인 관찰값들 중에서 모델이 True로 분류한 비율을 나타내며, 모델이 올바르게 관측치를 실제 class에 맞게 분류하였는지를 파악하여 모델의 완성도를 평가한다(Ballabio and Consonni, 2013).

정밀도(precision)은 positive로 분류된 값들 중에 정확하게 분류된 값들의 비율을 나타내며 모델의 분류성능을 가리킨다(Zhang et al., 2022).

F1 score는 재현율과 정밀도의 조화평균으로, 서로 반비례속성을 띄는 재현율과 정밀도가 같은 가중치를 가지게끔 만든 지표이다. F1 score는 데이터 분류군 간의 크기가 상이한 imbalanced 데이터를 이용한 분류모델을 평가지표로 사용된다(Jeni et al., 2013).

G_mean은 Geometric mean의 준말로 실제 positive를 제대로 분류할 확률(민감도, sensitivity)과 실제 negative를 제대로 분류할 확률(특이도, specificity)간의 기하평균으로, 마찬가지로 불균형 데이터를 이용한 분류모델을 평가할 때 사용된다.

변수중요척도(variable importance in projection)를 통한 변수선택법은 PLS모델에서 각 파장변수 X의 중요도를 보여주는 지표이다. Y의 총 변동 중 f번째 주성분에 기여한 PLS가중치를 합산하여 계산하며, 어떤 X변수가 가장 Y의 분산을 많이 설명하였는지를 점수 형태로 알려준다. VIP를 통해 점수화가 가능하기 때문에 각 파장변수들 간의 비교가 가능하고 평균점수는 1.0이기 때문에 보통 VIP = 1을 넘으면 중요한 변수로 평가된다(Farrés et al., 2015).

제 2 장 연 구 사

제 1 절 근적외선 품질분류

기존에 근적외선 파장검사를 이용한 연구(NIRS: Near Infrared Spectroscopy)에서는 주로 화학성분의 검출 확인을 목적으로 쓰여왔으나, 지속적으로 농업, 식품업 그리고 산림분야로 그 사용범위가 확장되었다. 특히 근적외선 파장을 이용한 종자 및 과실의 비파괴적 검사는 임업 및 농업 분야에서 최근에 각광을 받고 있다. 이는 검사 대상을 파괴하지 않고 품질을 측정할 수 있을 뿐만 아니라 품질 기준의 정량화를 통해 측정의 정확도를 높일 수 있고 기계화 및 자동화 공정을 통하여 빠른 시간 내에 대량으로 품질 검사가 가능하기 때문이다. 또한 종자의 품질과 직접적으로 연결된 발아율, 활력, 병충해 가해 여부를 판별하는 것은 특정 성분의 정량분석과 더불어 비파괴검사의 주요한 목적이 되어왔다.

국내의 경우, 근적외선 파장 검사를 이용한 종자의 각종 성분의 정량분석은 1990년대부터 꾸준히 연구되고 있지만, 종자의 품질에 관한 연구는 민태기 등에 의하여 박(*Lagenaria siceraria* Standl) 씨의 발아 여부에 대한 PLS2 모델 검정 연구(민태기와 강우식, 2003) 및 무 종자의 발아 여부에 대한 PLS2-DA 모델 검정 연구(Min and Kang, 2008), 무, 상추, 옥수수 등 종자에 대한 NIR 검정 및 PCA, PLS-DA을 통해 활력 여부 검정을 연구(Kang, 2009), 곰팡이 감염 감자와 온전한 감자의 분류에 대하여 PLS-DA 모델로 검정 연구(김대용 등, 2012), 그리고 수수와 수단그라스의 교잡종의 품종간 PLS-DA 검정 연구(이기원 등, 2020) 등 주로 농작물을 대상으로 한 연구가 이루어졌다. 하지만 농작물 외 국내외에서 산림용 종자 검증에 대해 근적외선 방법을 이용한 연구는 이루어진 바가 미비한 상황이다.

해외의 경우 근적외선 비파괴 검사는 종자의 보다 폭넓은 품질을 선별하

는 데 사용되어 왔다. 특히 국내에서는 연구된 바가 없는 종자 및 과실의 상품성이나 활력에 영향을 미치는 충해 여부 또한 근 20년간 다양한 수종에 걸쳐 꾸준히 연구가 되어온 바 있다. *Cordia africana*에 대한 과실 가해 여부를 NIR 파장을 이용한 PLS-DA 모델 검정 연구(Tigabu and Odén, 2002) 및 동아프리카 향나무(*Juniperus procera*) 종자에 대한 발아 여부 및 충해 여부를 VIS-NIR 파장대를 이용한 PLS-DA 모델 검정 연구(Tigabu et al., 2007), 대추(jujube)의 충해 여부에 대해 NIR 반사광, 투사광 별로 LRT-DA를 이용한 모델 검정 연구(Wang et al., 2010), 충해가 수분피해를 가하는 정도를 이용하여 밤의 충해 여부를 검정하는 LDA-DA모델 검정 연구(Moscetti et al., 2014), Chinese fir 건전종자와 충해여부를 포함한 충해종자를 선별하는 O2PLS-DA모델 검정 연구(Tigabu et al., 2019), 석류의 충해 가해 여부를 검정하는 PCA-DA모델 검정 연구 (Jamshidi, 2020), 내부가 부패한 사과에 대하여 SMOTE-CARS-SPA-LS-SVM 모델 검정 연구(Zhang et al., 2022)등 주로 농업용 과실수나 산지조림용 침엽수 종자를 대상으로 연구가 되어왔다. 그러나 기후변화 대비 산지조림용 수종인 상록활엽수종의 종자의 검사에 대한 연구는 거의 되어있지 않다.

제 2 절 붉가시나무

기후변화로 인하여 평균기온이 오르는 등의 국내 환경 변화에 따라 난대성 상록활엽수종의 수요 증가는 예상되어 왔다. 국내의 주요 난대성 상록활엽수종 중에서도 붉가시나무는 수요의 증가가 예상되어 국내에서 많은 연구의 대상이 되어왔다. 난대로 변한 환경에서의 탄소흡수량이 우수할 것으로 예측이 되며, 변화한 환경에서의 한반도 내 분포범위가 증가할 것으로 예상이 되기 때문이다,

붉가시나무의 탄소흡수능력은 2000년대 초부터 국내에서 꾸준히 연구되

어 왔다. 붉가시나무를 포함한 주요 난대상록수종에 대해 탄소배출계수를 이용한 탄소흡수량을 산정한 연구(Son et al., 2014), 완도 내 붉가시나무의 적지에 따른 탄소흡수량을 검정한 연구(Hwang et al., 2015), 붉가시나무의 상대생장식을 이용하여 탄소저장량 및 이산화탄소 흡수량을 계산한 뒤 소나무 등과 비교한 연구(Kim and Lee, 2017)가 이루어졌다. 해당 연구들에서는 국내 우점종인 소나무 및 타 수종들과의 탄소흡수능 비교가 주가 되어 왔으며, 붉가시나무의 탄소흡수능력은 국내 난대 상록활엽수종 중에서도 가장 우수하였다.

붉가시나무의 분포 및 잠재생육지역에 대한 예상에 관한 연구 또한 활발이 이뤄져 왔다. 1월 평균기온 및 최저기온, 강수량, 위도와 고도를 이용하여 난대성 상록활엽수종의 분포를 연구(구경아 등, 2001), 붉가시나무 등 난대성 상록활엽수종의 기후변화에 따른 분포도 및 자생북한계지를 연구(이정현과 최병희, 2010), 기후변화에 따른 최한월최저기온, 겨울철강수량을 이용하여 붉가시나무의 한반도 내륙 내 미래 분포 가능성을 연구(박선욱 등, 2016), 기후변화 및 분산 능력을 이용하여 붉가시나무의 잠재서식지를 연구(신만석 등, 2018) 등이 있었다.

제 3 장 연구 방법

제 1 절 연구재료

실험용 재료는 2021년 10월 13일 국립산림과학원 난대아열대산림연구소 붉가시나무 채종원(제주특별자치도 서귀포시 상효동 산1-5)(Figure 1) 에서 38가계를 대상으로 각 가계당 1-5개체를 선정하여, 각 개체당 약 5개씩 도토리 550립을 채취하였다. 채취 시 개화 및 결실이 이뤄지는 방향을 고려하여 4방위에서 종자를 가급적 골고루 채취하였다. 다만, 당해 붉가시나무 개화 및 결실 자체가 흉년이었기 때문에 가계별로 균등한 종자 채취는 어려웠다. 그로 인하여, 가계 별 최대 종자 개수는 18개, 최소 종자 개수는 1개, 가계 별 평균적인 종가 개수는 8.23개였다(Table 1). 그리고 크기를 고려하여 너비가 넓은 종자들을 채취하였다.

Table 1. Number of acorns collected from each family

Family	Acorns	Family	Acorns	Family	Acorns
기도원	1	선녀탕4	18	해송26	16
기도원1	17	선녀탕5	17	해송1	12
기도원2	10	선녀탕통합	9	해송2	5
기도원3	14	선녀탕2	3	해송3	23
남서교	19	수악교	8	해송서	6
동홍	17	수악교2	23	해송서1	9
동홍2-1	19	영주교	14	해송서13	15
동홍2-3	8	영천3	13	해송서4	14
동홍시험림	14	영천교정	16	호근다리	12
동홍통합1	25	완도수목원1-2	19	호근1-3	28
동홍1	14	한남	13	호근1-4	25
상효	11	해송1-1	17	화백절간	23
선녀탕1	11	해송18	12		

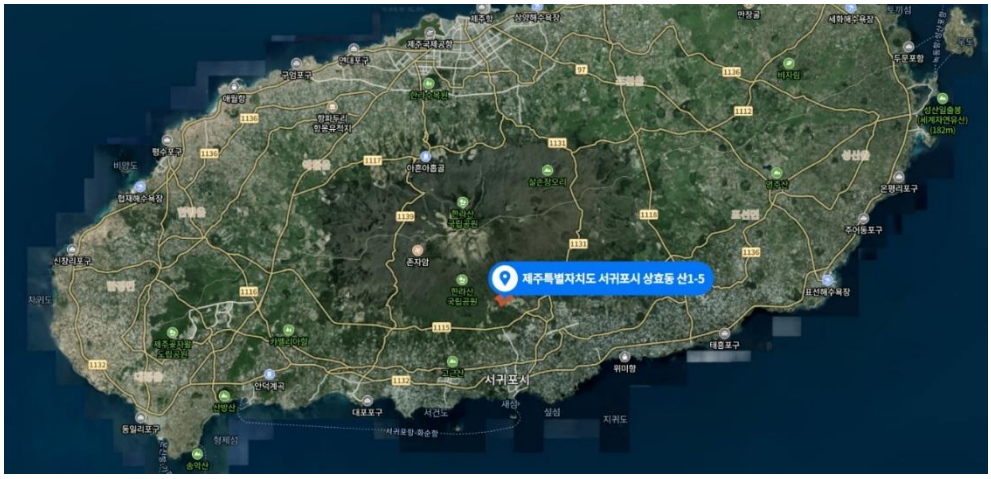

Figure 1. Location of the seedling seed orchard of *Quercus acuta* in Jeju

제 2 절 충해종자 분류

채취한 종자는 분광분석에 들어가기 전 엽록소를 분해 시키기 위하여 (Bonner, 1987) 약 4개월간 5도의 냉장환경에서 보관하였다. 이후 종자를 분류를 하였다. 도토리 종자를 측정하여 종자분석을 진행하였고, 이후 시각적인 판단에 의하여 분류하였다. 종자분석의 경우, 종자 폭, 종자길이, 종자 무게 등을 측정하였다. 종자 분류의 경우, 충해 여부(벌레가 먹었는가, 먹지 않았는가), 종피의 변색 여부, 종피 머리 부분의 부패 여부 등이었다. 이를 이용하여 종자의 건전 정도를 분간하였다.

충해 여부의 경우, 사전연구에서 충해에 관련하여 분류에 사용한 종자가 해자들의 가해여부를 참고로 하였다(Dorsey et al., 1962; 최광식, 2006). 도토리 바구미 애벌레의 존재가 확인되거나(Figure 2. left), 산란구멍(Figure 2. middle) 및 탈출공(Figure 2. right)이 확인되었을 경우, 충해종자로 분류하였다. 종자를 시각(visual inspection)으로 분류한 결과 총 368립의 건전한 종자, 182립의 충해종자로 구분할 수 있었다.

Figure 2. Picture of the caterpillar of *Curculio dentipes* (left), acorn with a small hole (middle) and acorn with an exit hole (left)

제 3 절 종자 전처리

근적외선 분광분석 측정 시, 간섭요소로 작용하는 수분과 온도조건을 배제하기 위해 다음과 같은 전처리 과정을 거쳤다. 온도의 경우, 냉장환경에서 바로 꺼낸 도토리의 온도를 실온과 비슷하게 맞춰 주기 위하여 도토리들을 25도의 항온수조(Figure 3)에 10분간 방치했다(Takahashi et al., 2011). 이 때 이러한 온도처리에 의해서 도토리 내부 전분의 deterioration rate가 영향을 받는 범위는 최소 24시간 내지 수일간의 온도변화에 의해 영향을 받기 때문에 10분간의 실온처리는 발아율 및 종자 품질에 영향을 주지 않을 것이라는 판단을 하였다(Khan and Habib, 2015).

Figure 3. Water bath treatment for temperature management

제 4 절 NIR 파장 측정

NIR 파장 측정을 위한 근적외선 분광분석기는 Oceanview Flame S-Vis-Nir를 사용하였다. 근적외선 광원 제공을 위한 적분구는 Oceanoptics사의 integrating sphere for reflectance(ISP-REF)를 사용하였으며, 분광분석기와 적분구를 연결하는 광섬유의 경우, Oceanoptics사의 QP400~1-Vis-Nir(슬릿 사이즈 25um, 그레이팅타입 600)를 사용하였다. Ocean Insight사에서 제공하는 resolution calculator를 사용하여 계산한 resolution은 약 1.39cm^{-1}이었다. 분광분석 시, reflectance 계산에 필요한 white reference 및 dark reference의 경우, Oceanoptics 사의 diffuse reflectance standard, spectralon (WS-1SL)를 사용하였다. 조사 파장대는 가시광선 영역대와 근적외선 영역대의 일부를 포함한 400nm-1000nm로, 해당 실험기기로 조사가 가능한 최대 범위였다.

적분구를 통해 만들어진 아날로그 형태의 반사광은 NIR spectrometer를 통해 디지털 신호로 변환된다. 이렇게 측정된 디지털 신호는 반사도(Reflection, R(%))라 하며 다음과 같은 공식에 의해 계산하였다(수식{1}).

$$\boldsymbol{R_n = \frac{R_s - R_{back}}{R_{ref} - R_{back}} \times 100}(\%) \qquad \{1\}$$

n번째 파장에서의 반사도(R_n,%)는 샘플의 반사광(R_s)으로부터 광원이 없는 상태에서 측정한 background reflection(R_{back})을 뺀 수치에 대한 표준물질의 반사광(R_{ref})에서 background reflection을 뺀 수치의 비율로 계산할 수 있다. 반사도는 상대적인 수치이기 때문에 아래 수식을 통해 흡광도(Absorbance, A_n, AU) 스펙트럼으로 변환하였다(수식{2}). 로그 수식의 밑의 경우, 실험 대상이 종자, 즉 생체이므로 자연로그를 사용하였다.

$$A_n = \log_e(1/R_n) \qquad \{ 2 \}$$

측정된 디지털 데이터는 Ocean Insight사의 Oceanview spectroscopy software 2.0.8를 이용하여 반사도를 계산한 뒤, 프로그램에 탑재된 보간법 기능을 이용하여 파장분석을 1nm 구간별로, 즉 400nm-1000nm 구간대에서 총 601개의 파장변수로 전환되었다. 샘플의 측정은 integrating sphere의 1cm*1cm 크기의 측정구멍 바로 위에 위치하여 종자의 한 단면을 측정이 가능하도록 하였다(Figure 4). 도토리의 모든 단면을 고려하기 위하여 90도마다 10번씩 측정하여 총 40번을 측정한 뒤 평균 낸 값을 spectra로 사용하였다. NIR 측정 시 샘플의 특성정보를 반영치 않는 불필요한 specular reflection을 제거하고 diffusively reflected light만을 측정하기 위하여 적분구의 세팅을 exclude로 조정하였다.

NIR 측정 시 발생할 수 있는 오류들을 막기 위해 여러 조치를 수행하였다. 우선 프로그램에 내장된 여러 기능들을 사용하였다. STA(scans to average)는 15, BW는 5로 설정하였다. STA는 몇 번을 측정한 값을 평균치로 할 인지를 의미하는 수치로 분광분석 과정에서 나오는 오류들을 줄여준다. Non-linear correction 기능은 integrating sphere 등의 기기에서 발생하는 오류를 잡아준다(Nehir et al., 2019). 그리고 도토리 근적외선 파장 측정 시 주변 광원으로 인한 간섭을 막기 위하여 검은 박스 형태의 차단막(Lid)(Figure 4, ①)을 만들어 사용하였다. 도토리에서 반사된 빛 이외에 외부의 불필요한 광원의 유입은 조사결과를 방해할 수 있기 때문에 종자를 적분구 위에 올려놓고 그 위에 차단막을 덮어서 어두운 환경을 조성했던 사전연구결과를 참고하였다(Daneshvar et al., 2015).

또한 수십회 측정 이후, ‘instrumental drift’ 라고 불리는 장비 측정오류가 생기는 점을(Farhadi et al., 2016) 고려하여 종자 5개 측정시마다 light

reference와 darkness reference를 갱신하였다. 이렇게 측정 이후 생기는 오류는 detector에 light saturation이라는 현상이 발생하기 때문이다 (Moscetti et al., 2013). 이러한 오류를 고려하여 측정 시작 시 적분구를 30분 이상 미리 가동하여 측정에 오류가 생기는 것을 방지하였다. 최종적으로 도토리에 대한 종자 측정은 다음과 같은 배치를 통해 이뤄졌다(Figure 4).

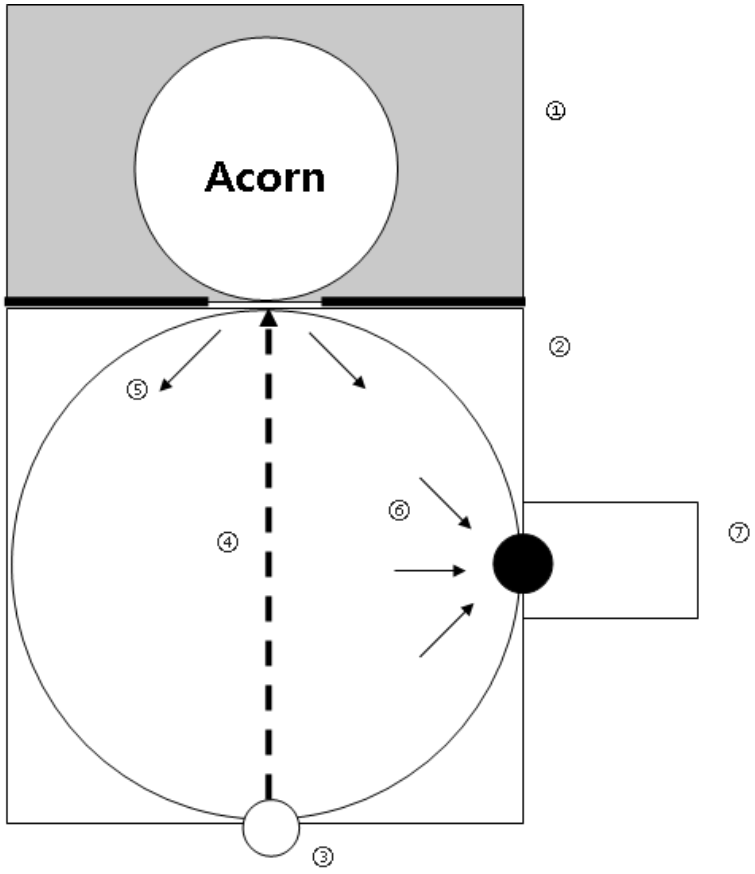

Figure 4. Acorn spectroscopy using integrating sphere. ①: Lip, ②: Integrating sphere, ③: Light source, ④: Light from the source, ⑤: Sample reflected light, ⑥: Reflected light processed by scattered light exception within the sphere, and ⑦: Optical fiber –moves the reflected light from the sphere to a detector

제 5 절 데이터 분석

제 1 항 데이터 전처리

전처리 방법의 경우, 모든 샘플에 대해 공통적으로 평균중심화(mean-centering)을 진행하였다. 이후 표준정규화 혹은 표준정규변동(Standard Normal Variate), 1st, Savitsky-Golay filter, Multiplicative Scatter Correction 등을 수행하였다.

표준정규화는 원 파장값에서 파장 평균을 차감한 후 표준편차로 나눔으로써, 파장의 평균이 0, 표준편차가 1이 되게끔 유도하는 전처리이다(수식 {3}). MSC(multiplicative scatter correction)는 파장에서 상가적 효과, 파장변수들 간의 상호작용으로 인한 효과를 제거함으로써 원파장값을 찾는데 도움을 주는 전처리이다(수식{4}). 표준정규화 및 MSC를 이용하여 광산란이 발생시키는 간섭효과들을 제거하였다(Næs et al., 2002).

$$\boldsymbol{x_{snv}^{k} = \frac{x_{ori}^{k} - \sum x_{ori}^{k} / k}{\sigma_{x_{ori}^{k}}}} \qquad \{3\}$$

(x_{snv}^{k}: 표준정규화를 거친 k번째 파장의 파장변수, x_{ori}^{k}: 전처리를 거치지 않은 원래 파장변수)

$$\boldsymbol{x_{nk} = a_n + b_n \overline{x_k} + e_{nk}\ (n = 1 \sim N, k = 1 \sim K)}$$
$$\boldsymbol{x'_{nk} = (x_{nk} - \widehat{a_n}) / \widehat{b_n}} \qquad \{4\}$$

(x_{nk}: 간섭효과를 받지 않은 n번 샘플의 k번째 파장의 파장변수, x'_{nk}:MSC 처리과정을 거친 후의 k번째 파장의 파장변수,

a_n: 상가적 효과(additive effect), b_n: 상호작용(multiplicative effect), e_{nk}: 상가적, multiplicative effect로는 설명이 불가능한 영역)

1차 미분의 경우, 각각 기준값 오류(baseline)만을 지우거나, 기준값 오류뿐만 아니라 선형 추세(linear trend)를 제거하는 효과를 가지고 있다. 미분처리는 다음과 같은 공식을 사용하였다(수식{5}).

$$\begin{aligned} x'_n &= x_n - x_{n-1} \\ x''_n &= x'_n - x'_{n-1} \end{aligned} \qquad \{5\}$$

SG(Savitsky-Golay fileter)는 미분을 이용한 smoothing, 즉 잡음제거방법의 일종이다. SG전처리를 이용하여 추정다항식을 구성하는데 사용하고자 하는 파장의 개수(window size)와 다항식의 차수(the degree of the fitted polynomial)를 정하여 다항식을 계산한 뒤, 사용된 파장대의 가장 중심에 있는 파장값에서의 n차 미분값을 사용하는 방식으로 진행하였다. 이 과정에서 기준값 효과(baseline effect)를 제거하였으며 흡광지점에서의 흡수도를 더욱 증폭시켰다.

제 2 항 다변량 분석기법 적용

근적외선 분광분석 시 관측대상물질을 구성하는 분자들로 인해 발생하는 흡광지점 간에 중첩이 발생하기 때문에 파장변수들간 심한 상관관계를 제거할 수 있는 다변량 분석기법을 사용해야 한다(정회일과 김효진, 2000). 다변량 분석기법은 PCA, PLS-DA와 OPLS-DA를 사용하였다. Simca-P 14.1(umetrics)를 이용하여 분석을 하였다.

데이터를 통해 만든 건전-충해 분류 모델의 유효성을 검증하기 위해 calibration set, validation set을 구성하여 cross validation을 위한 구분을 하였다. Calibration set과 validation set의 크기는 각각 366립, 184립으로 지정하였으며, 관찰된 자료의 각 set 배치는 R에 내장되어 있는 sample, rep 등의 함수를 이용하여 랜덤으로 배치하였다. Calibration set에는 건전종자 246립, 충해종자 120립, validation set에는 건전종자 123립, 충해종자 61립을 배정하였다.

본격적으로 분석에 들어가기 전에 전처리를 거치지 않은 원래 파장값들과 전처리를 거친 데이터에 대해서 이상치 제거 작업을 하였다. 이상치 제거 시 사용한 방법은 다변량 데이터 분석 시 많이 사용하는 Hotelling's T^2 통계량을 사용하였다. 호텔링의 T^2 통계량은 일반적인 Student's T 분포를 다변량 분석에 적용할 수 있게끔 바꾼 방법이다. PCA를 통해 계산한 1,2번째 주성분을 이용하여 95% 신뢰구간을 구한 뒤, 5%의 1종 오류 영역에 속한 관찰값들을 이상치로 지정하여 제거하는 방법이다. 이상치 제거는 calibration set에서만 시행하였다(Wise et al., 2006).

구체적인 분석 방법으로는, NIR 파장 데이터를 X라 하고, 종자 품질 분류를 통해 얻은 해충 가해 여부를 Y라 하였을 때, X는 601개의 파장변수로 구성된 550개의 파장데이터(550 * 601), Y는 -1, 1로 구성된 가변수 형태의

데이터로 만들어진다. 파장데이터에 대해서는 1,2차 미분, MSC, Savitsky-Golay filter, SNV 등의 데이터 전처리를 이용하여 NIR 측정할 시 발생하는 오류를 제거하였다. 이후 calibration set, validation set으로 랜덤으로 368, 182개의 파장값을 분리하였다. 다변량 분석을 하기 전에 파장변수들의 평균값을 0으로 만들기 위하여 mean-centering을 시행하였다. Calibration set에 대하여 각 전처리와 PLS-DA, OPLS-DA을 조합하였다. 만들어진 calibration 모델에 대해서는 7-fold cross-validation을 시행하였다.

제 3 항 모델 성능 평가지표

만들어진 분류 모델에 대해서는 어떤 조합으로 만들어진 모델이 붉가시나무 종자의 종자 가해 여부 판별에 가장 적합한지를 비교하였다. 또한 400~1000nm 구간대는 400~780nm까지의 가시광선 영역, 780~1000nm의 근적외선 영역으로 이루어져 있기 때문에, 400~1000nm의 Vis-Nir 영역, 400~780nm의 가시광선, 780~1000nm의 근적외선 영역대별로 파장변수를 나눠서 각 영역대별 모델성능 또한 비교하였다. 모델성능은 classification evaluation index를 이용하여 비교하였다.

Classification에 쓰이는 기초적인 지표들로는 TP, TN, FP, FN이 있다. TP는 True Positive의 약어로 모델에 의해 1 또는 True로 분류된 관찰값이 실제로 true인 경우의 수를 의미한다. FP는 false negative의 약어로 모델에 의한 분류 및 실제 분류 둘 다 positive인 경우의 수를 의미한다. FP는 false positive의 약어이며, 모델에 의한 분류는 positive이나 실제로는 negative인 경우의 수를 의미한다. 마지막으로 TN은 true negative의 약어로, 모델이 실제로 Negative인 관찰값을 제대로 Negative로 분류한 경우의 수를 의미한다. 이러한 4가지 성분들을 기초로 하여 분류모델의 성능을 평가하는 분류지표들을 만들 수 있다(Table 2). 이후 변수중요척도(수식 {6})를 이용하여 각 모델별로 어떠한 파장변수들이 영향을 많이 주는지를 판별하였다.

$$\boldsymbol{VIP_k} = \sqrt{\frac{\sum_{f=1}^{F} \boldsymbol{w_{kf}^2 SSY_f}}{\boldsymbol{SSY_{total}} * \boldsymbol{F}} * \boldsymbol{K}} \qquad \{6\}$$

(VIP_k: k번째 파장변수X의 VIP, w_{kf}: k번째 파장변수와 f번째 주성분의 PLS가중치, SSY_k: f번째 주성분 내에서의 Y의 설명가능분산, SSY_{total}: 모든 주성분 내 Y의 설명가능 분산, F: 주성분의 숫자)

Table 2. Evaluation of accuracy and precision

Accuracy	(TP+ TN)/(TP+ TN+ FP+ FN)
Recall	(TP)/(TP+ FN)
Precision	(TP)/(TP+ FP)
F1_Score	(2*Precision*Recall)/(Precision+ Reccall)
G_mean	$\sqrt{\frac{TP}{TP+FN} * \frac{TN}{TN+FP}}$
TP	Actual positive-pred postive
FP	Actual negative-pred positive
FN	Actual positive-pred negative
TN	Actual negative-pred negative

제 4 장 연 구 결 과

제 1 절 NIR 파장의 형태 분석

각 종자당 사방위로 10번씩, 총 40개의 파장을 평균값을 낸 결과, 550개의 근적외선 파장들을 확보할 수 있었다. 그 결과값은 다음과 같다(Figure 5). 전반적으로 가시광선 영역인 400nm-780nm에서 높은 수치를 보이고 있고, 근적외선 영역대인 780~1000nm 구간대에서는 낮은 흡수율을 보이고 있었다. 또한 680nm 및 938nm에서 각각 뚜렷하거나 느슨한 흡광지점을 보여주었다.

사전에 분류된 바에 따라 건전종자, 충해종자 별로 평균값을 낸 NIR 측정값은 다음과 같았다(Figure 6). 600nm 이후로 건전종자, 충해종자 간의 구분이 나타나기 시작하지만, 적외선 영역인 780nm 이후로 더 명확히 평균값들 간에 차이가 발생하였다. 가계에 따른 구간대의 차이는 거의 존재하지 않았다($p<0.05$).

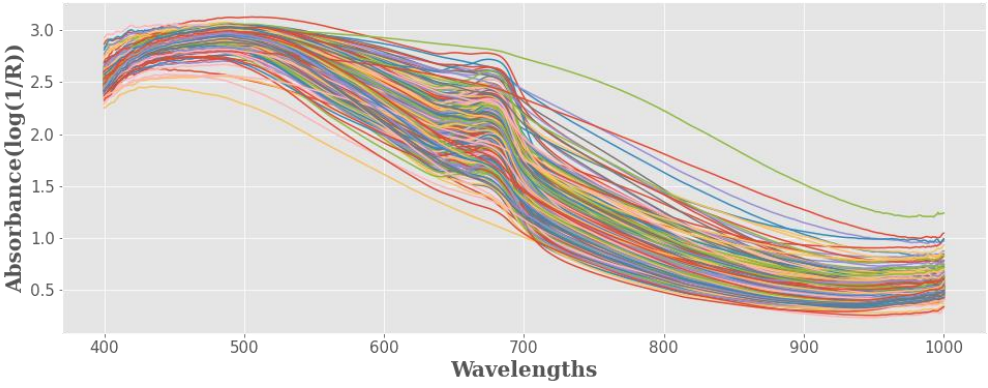

Figure 5. Whole VIS-NIR raw spectra from 400 to 1000nm for acorns of *Quercus acuta*

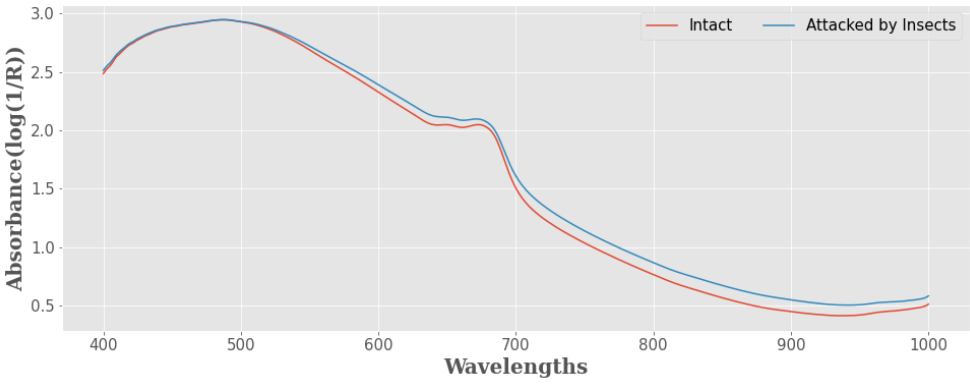

Figure 6. Average of wavelength between sound and damage acorns from 400 to 1000nm

제 2 절 충해종자 검정 모델 간 비교

우선 전처리 여부가 실제 분류에 영향을 주는지를 알아보기 위하여 원데이터를 이용하여 분석하였다. 원자료의 데이터의 분포 상태를 파악하기 위하여 400~1000nm, 400~780nm, 780~1000nm 구간대의 calibration set에 95% 신뢰구간으로 주성분 분석(PCA)을 처리하였다. 그 결과 score plot 상에서 건전종자, 충해종자 분류군 간 어느정도 클러스터링이 발생하는 것을 확인하였다. 780~1000nm의 파장대의 제1주성분(96.7%), 제2주성분(3.94%)으로 구성된 score plot 상에서 건전종자군들 간의 명확한 클러스터링을 확인하였고 충해종자군들의 경우 느슨한 군집을 형성하며 넓게 퍼져 있는 것을 확인하였다(Figure 7).

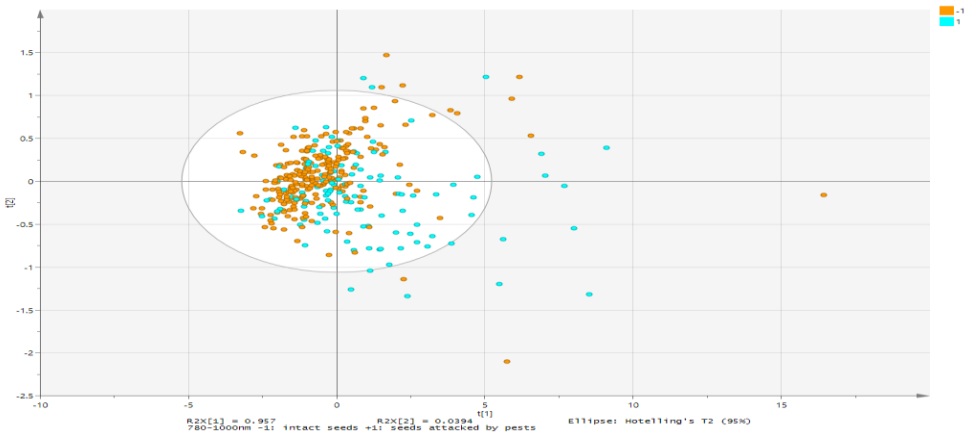

Figure 7. Principal component analysis for sound and damaged acorns in wave range from 780 to 1000nm

실제로 파장구간대에 따라 다변량 분석 모델 성능이 차이가 나는지를 확인하기 위하여 전처리가 안된 calibration set을 대상으로 400~1000nm, 400~780nm, 780~1000nm을 나눠서 분석하였다(Table 3). 모든 구간대는 재현율이 95% 이상으로 건전종자의 분류에 대해서는 높은 분별력을 보여주었다. 특히 400~780nm 구간대에서는 PLS-DA, OPLS-DA 각각 98.37%, 99.19%의 높은 재현율을 보여주었다. 그러나 충해종자의 분류 또한 반영하였을 때는 다른 양상을 보여주었다.

400~1000nm 내에서는, PLS-DA, OPLS-DA 각각 85.87%, 77.17%의 정확도를 보여주었다. 근적외선 영역인 780~1000nm의 경우, PLS, OPLS-DA 둘 다 85.87%로 동일한 정확도를 보여주었다. 반면 가시광선 영역만을 분석한 400~780nm 구간대의 경우, 정확도가 67-69%로 다른 두 구간대에 비해 모델 성능이 낮았다.

G mean의 경우, 이러한 차이가 더 두드러졌다. 400~780nm대에서의 G mean은 PLS, OPLS-DA 각각 33.60%, 22.09%로 65-80%인 다른 구간대에 비하여 매우 낮은 수치를 보여주었다.

Table 3. Evaluation index for raw spectra

Preprocessed	Raw spectra					
Wavelength	400~1000nm		400~780nm		780~1000nm	
Model name	RawWP	RawWO	RawVisP	RawVisO	RawInfP	RawInfO
Analysis	PLS-DA	OPLS-DA	PLS-DA	OPLS-DA	PLS-DA	OPLS-DA
Components	12	1+3	2	1+0	7	1+5
Accuracy	0.8587	0.7717	0.6957	0.6793	0.8587	0.8587
Recall	0.9512	0.9350	0.9837	0.9919	0.9593	0.9512
Precision	0.8540	0.7718	0.6914	0.6778	0.8489	0.8540
F1_Score	0.9000	0.8456	0.8121	0.8053	0.9008	0.9000
G_Mean	0.7996	0.6433	0.3360	0.2209	0.7931	0.7996

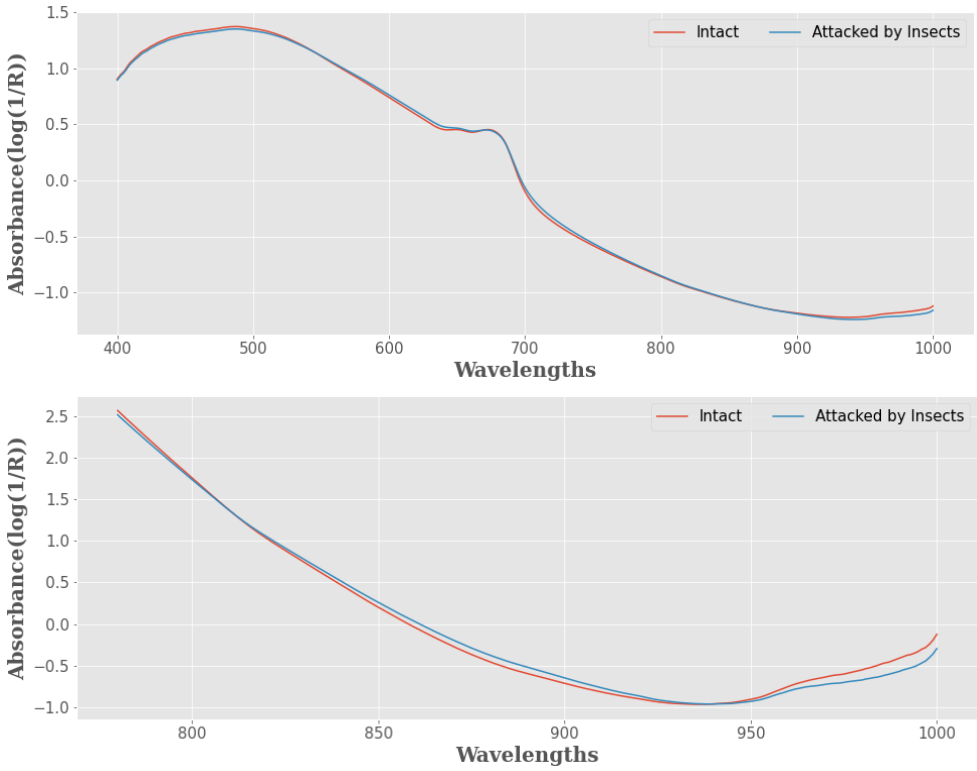

Figure 8. Spectra processed with standard normal variate 400~1000nm (upper) and 780~1000nm (lower)

원자료를 이용했을 시와 통계적 전처리를 거친 데이터를 이용해 분석했을 때 분석모델의 차이를 비교하기 위하여, calibration set에 대하여 여러 전처리를 가하였다. 우선 SNV(standard normalized variate)를 처리한 결과, 원래 파장값들에 비해 variation은 줄어들었고 680nm 및 940nm 부근의 흡광지점은 여전히 보였으나, 평균을 차감하는 절차에 의하여 여러 구간대별로 파장들이 교차하였다(Figure 8. upper). SNV 전처리 하에서의 건전 종자 평균, 충해 종자 평균 스펙트럼의 경우, 이러한 교차로 인하여 잘 구분되지 않는 모습을 보였으나, 780nm-1000nm의 근적외선 구간만을 따로 전처리를 한 경우, 950nm 이상에서는 큰 차이로 갈라지는 모습을 보여주었다(Figure 8. lower). 실제로 다변량 분석기법 모델을 구축하였을 때 가시광선 영역대를 포함한 경우는 정확도, F1 score, G mean 각각 80%, 12.75%로 낮았으

나 근적외선 영역대만을 대상으로 모델을 구축한 경우 87.69%, 75.98%으로 더 높았다(Table 4).

Table 4. Evaluation index for standard normal variate processed spectra

Prepro-cessed	Standard Normal Variate (SNV)			
Wavelength	400~1000		780~1000	
Model name	SVWP	SVWO	SVInfP	SVInfO
Analysis	PLS-DA	OPLS-DA	PLS-DA	OPLS-DA
Compo-nents	1	1+0	7	1+6
Accuracy	0.6685	0.6685	0.8261	0.8261
Recall	0.9919	0.9919	0.9268	0.9268
Precision	0.6703	0.6703	0.8321	0.8321
F1_Score	0.8000	0.8000	0.8769	0.8769
G_Mean	0.1275	0.1275	0.7598	0.7598

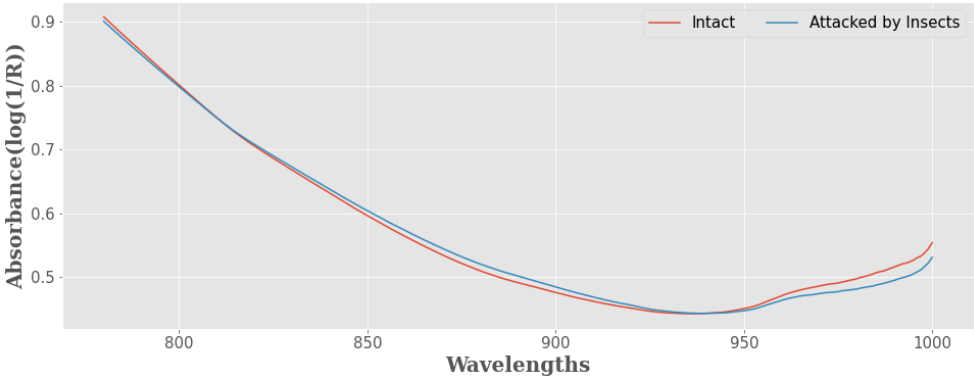

Figure 9. Spectra processed with multiplicative scatter correction at 780~1000nm

Multiplicative Scatter Correction(MSC)를 전처리로 사용한 데이터의 경우, SNV 전처리 데이터와 비슷한 양상을 보였다. 전 파장 및 근적외선 파장대를 대상으로 한 경우 분류군간 차이가 발생하는 흡광지점을 찾기 힘들었으나, 780~1000nm 구간대에서는 여러 구간대에 걸쳐 분류군간 차이가 발생하였다(Figure 9). 실제로 400~1000nm 구간을 이용한 모델 성능의 경우, PLS, OPLS-DA 둘 다 66.85%, 12.75%의 낮은 정확도와 G mean을 보여주며 충해종자의 분류가 되기 힘들었다. 반면 780~1000nm를 대상으로 한 경우 PLS, OPLS-DA 둘다 85.63%의 정확도를 보여주었다(Table 5).

Table 5. Evaluation index for multiplicative scatter correction processed spectra

Prepro-cessed	Multiplicative Scatter Correction (MSC)			
Wavelength	400~1000		780~1000	
Model name	MWP	MWO	MInfP	MInfO
Analysis	PLS-DA	OPLS-DA	PLS-DA	OPLS-DA
Components	1	1+0	6	1+5
Accuracy	0.6685	0.6685	0.8478	0.8478
Recall	0.9919	0.9919	0.9350	0.9350
Precision	0.6703	0.6703	0.8519	0.8519
F1_Score	0.8000	0.8000	0.8915	0.8915
G_Mean	0.1275	0.1275	0.7927	0.7927

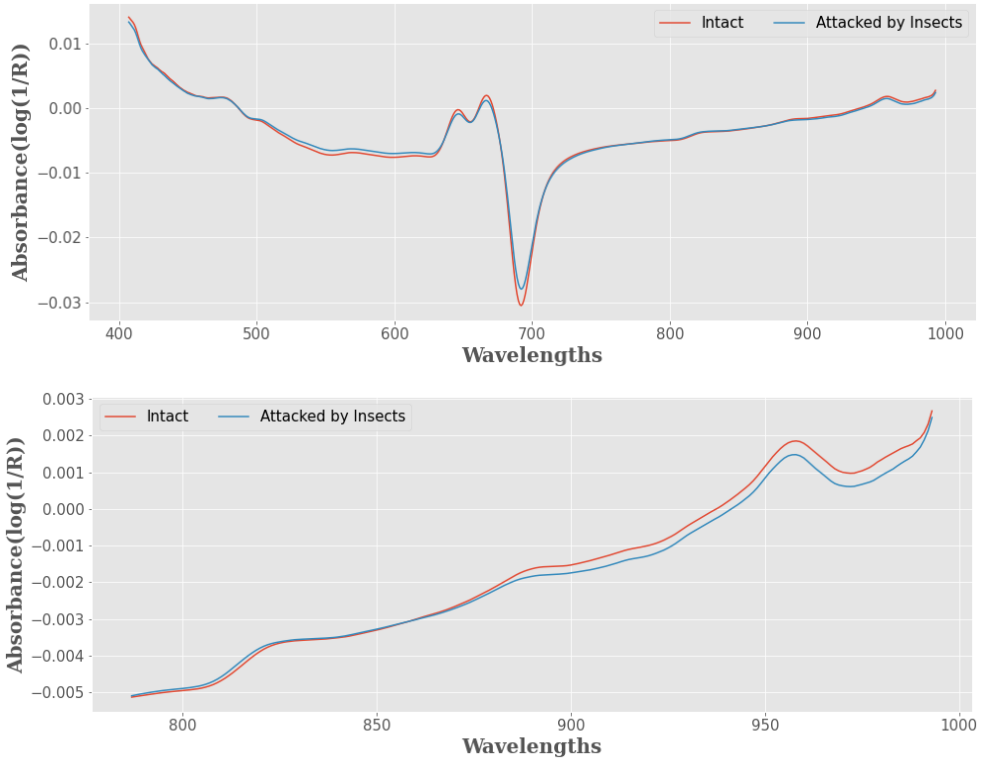

Figure 10. Spectra processed with 1st derivative 400~1000nm (upper) and 780~1000nm (lower)

1차 미분 전처리를 한 경우, 400~1000nm의 전 파장에 대해서는 680nm, 690nm, 980nm부근에서 흡광지점이 발생하는 것을 볼 수 있었다. 그러나 전반적으로는 분류군들 간에 분리되는 경향을 찾기가 힘들었다(Figure 10. upper). 반면 780~1000nm만을 대상으로 전처리를 하였을 때, 960nm 부근에서 강한 흡광지점을 확인하였으며 850nm 이후로 건전종자와 가해종자 간의 차이가 발생하는 것을 볼 수 있었다(Figure 10. lower). 전처리된 데이터들을 대상으로 다변량 분석을 하였을 때 400~1000nm 및 780~1000nm 구간에서의 모델의 정확도는 82~84%로 비슷하였으며, 각 파장대 내의 PLS, OPLS-DA별 모델성능 또한 F1 score 및 G mean이 87~88%, 75~78%로 거의 차이가 나지 않았다(Table 6).

Table 6. Evaluation index for 1st derivative processed spectra

Preprocessed	1st derivative			
Wavelength	400~1000		780~1000	
Model name	1DWP	1DWO	1DInfP	1DInfO
Analysis	PLS-DA	OPLS-DA	PLS-DA	OPLS-DA
Components	5	1+8	6	1+7
Accuracy	0.8261	0.8424	0.8424	0.8207
Recall	0.9187	0.9268	0.9268	0.9187
Precision	0.8370	0.8507	0.8507	0.8309
F1_Score	0.8760	0.8872	0.8872	0.8726
G_Mean	0.7664	0.7893	0.7893	0.7565

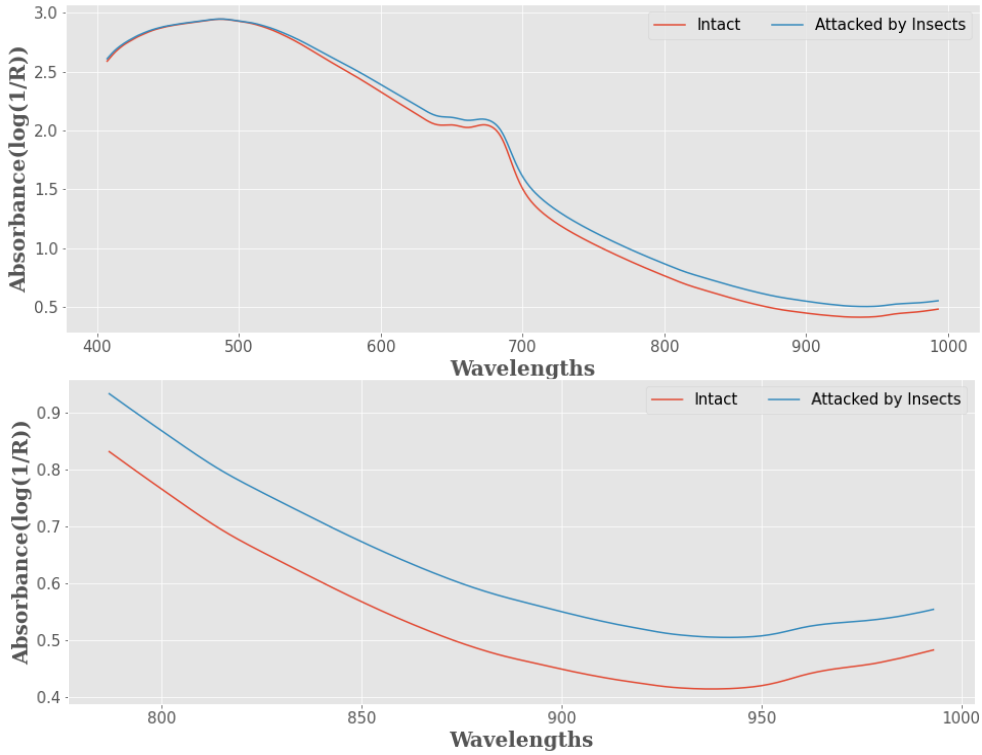

Figure 11. Spectra processed with Savitsky-Golay filter, 400~1000nm (upper) and 780~1000nm (lower)

마지막으로 SG(Savitsky-Golay filter) 전처리를 적용한 경우, 400~1000nm의 전 파장을 대상으로 한 경우, 원파장과 마찬가지로 680nm 및 900~1000nm 구간에 걸친 흡광지점이 나타났으며, 전 파장에 걸쳐 분류군 간의 차이가 발생하는 양상을 보였다(Figure 11. upper). 780~1000nm의 근적외선 구간만을 대상으로 전처리를 한 경우에는 900~1000nm의 흡광지점을 바탕으로 분류군들 간의 차이가 더 벌어지는 양상을 보였다(Figure 11. lower). 전처리된 데이터를 이용하여 다변량 분석을 하였을 때는 400~1000nm에 대해 PLS-DA 분석을 하여 얻어낸 모델이 정확도가 86.41%로 780~1000nm을 대상으로 한 다변량 분석과 비슷하거나 더 높은 분별력을 보였다(Table 7).

Table 7. Evaluation index for Savitsky-Golay filter processed spectra

Prepro-cessed	Savitsky-Golay filter (SG)			
Wavelength	400~1000		780~1000	
Model name	SGWP	SGWO	SGInfP	SGInfO
Analysis	PLS-DA	OPLS-DA	PLS-DA	OPLS-DA
Components	11	1+3	7	1+6
Accuracy	0.8641	0.7717	0.8315	0.8315
Recall	0.9512	0.9350	0.9350	0.9350
Precision	0.8603	0.7718	0.8333	0.8333
F1_Score	0.9035	0.8456	0.8812	0.8812
G_Mean	0.8093	0.6433	0.7632	0.7632

각 전처리와 다변량 분석 별 조합을 통한 모델들 중에서는 400~1000nm 파장대에 대해 SG로 전처리를 한 데이터에 PLS-DA를 적용한 SGWP모델이 86.41%로 가장 높았으며, 정밀도, F1 score, G mean 또한 각각 86.03%, 90.35%, 80.93%로 건전종자, 충해종자 분류에 있어 가장 우수한 분류성능을 보여주었다. 그러나 전처리를 가하지 않은 모델들의 경우 역시 정확도, F1 score, G mean이 89~90%, 79~80%로 SG 전처리와 비교하여 거의 차이가 나지 않거나 대부분의 전처리 방법에 비하여 높은 분류 성능을 보였다 (Figure 12).

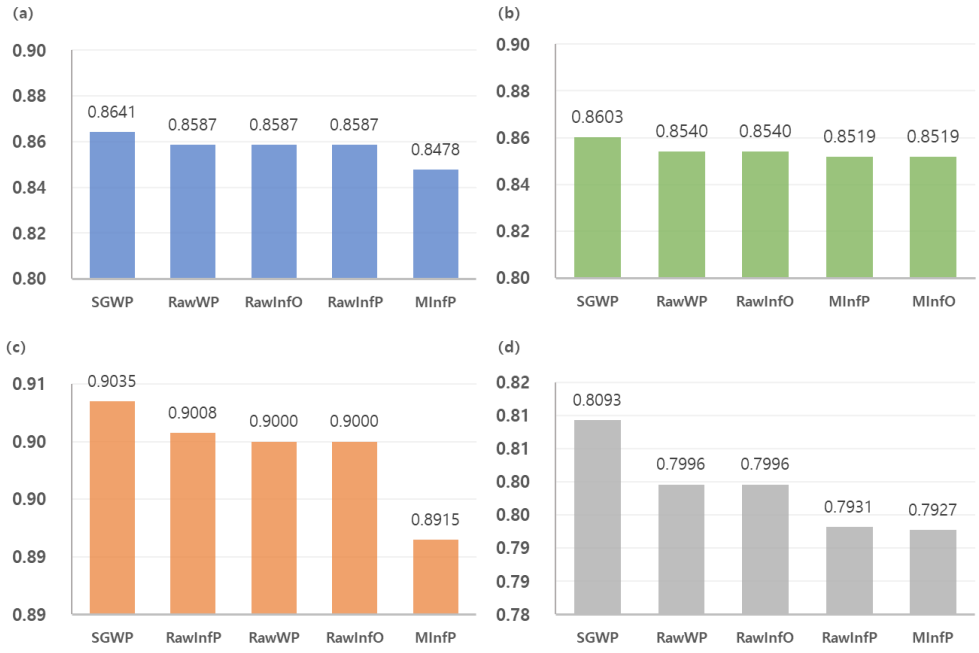

Figure 12. Top 5 evaluated combinations of analyzation methods, accuracy (a), precision (b), f1 score (c), G mean (d), SG: Savitsky-Golay, Raw: No preprocessing, W: 400~1000nm, Inf: 780~1000nm, P: PLS-DA, O: OPLS-DA

해당 모델의 데이터 분류 성능을 시각적으로 판별하기 위해 score scatter plot을 확인한 결과는 다음과 같았다. 제1 주성분(77.7%의 분산을 설명), 제2 주성분(13%의 분산을 설명)으로 구성된 scatter plot만으로는 확연히 분리되는 경향을 확인하기는 힘들었으나 건전종자, 충해종자군 별로 군집이 형성되는 것은 확인할 수 있었다(Figure 13).

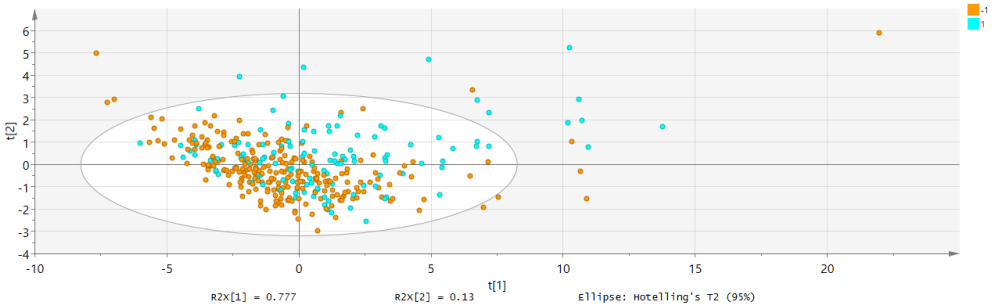

Figure 13. 400~1000nm Savitsky-Golay filter (SG) processed model's PC score plot, t1: PLS principal component that take the largest part of the explainable variation, t^2: the 2nd largest PC

PLS 모델의 실제 모델의 샘플 분류지표로 쓰이는 y-predictability는 0.5로 설정해둔 기준선(threshold)을 중심으로 하여 observation들을 0,1으로 나누게 된다. SGWP 모델의 경우, 건전종자를 실제 건전종자 116개 중 111개를 제대로 분별하였고, 충해종자 58개중 41개를 제대로 분별하였다(Figure 14).

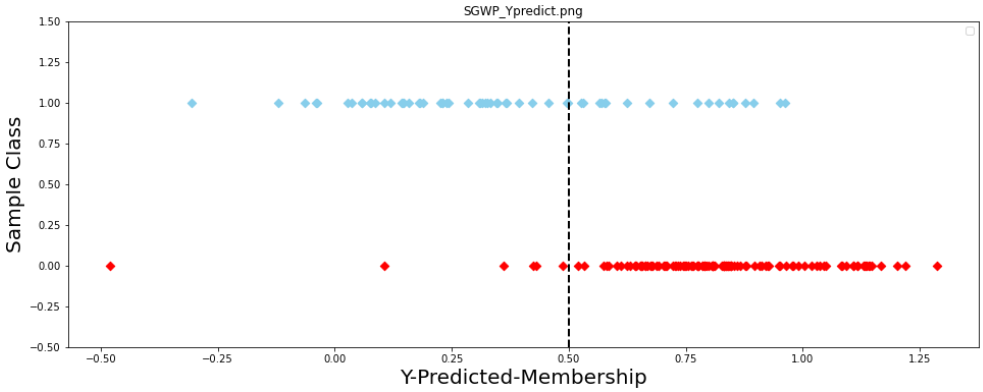

Figure 14. Y-predicted plot, it shows how y predicted values from PLS regression are set to classes by a certain threshold (Y = 0.5). Sample classes of 0,1 mean original classification of the samples. Blue dots: Sound acorns, red dots: damaged acorns

모든 변수를 사용하였을 경우, 실제 분광분석의 분류 사용시 연산의 문제가 발생할 수 있기 때문에, 실제 모델에 있어 어떤 파장변수를 사용하는 것이 바람직한 지, 차후 어떤 파장을 선택하는 것이 적절한지를 파악하기 위하여 VIP를 통한 변수 선택을 하였다. VIP=1을 넘는 파장변수는 161개였으며, 주로 650~750nm 사이의 가시광선 영역대와 400nm 부근, 그리고 950nm 이상의 변수들이 이에 해당하였다(Figure 15). Regression coefficient 또한 마찬가지로 650~750nm에서 강한 peak가 두드러지고, 400nm, 950nm 부근 순으로 다른 peak가 나타나는 것을 확인하였다(Figure 16).

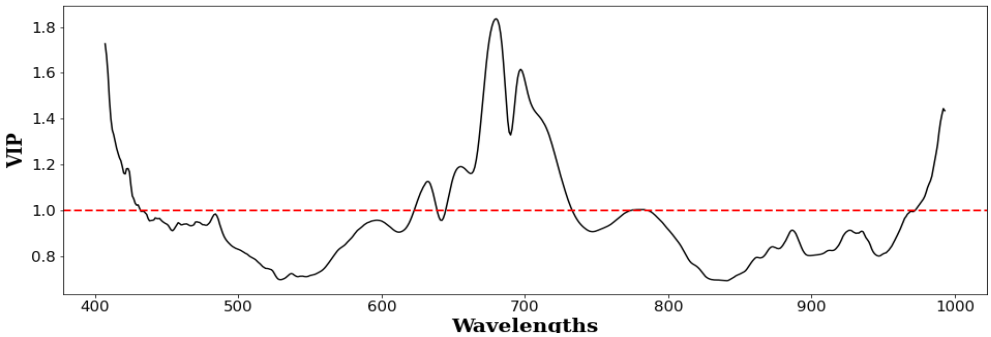

Figure 15. Variable Important for the Projection: A normalized index that explains which spectra in the model were responsible for Y-predicted-value decision. Mean squared VIP in every wavelength is always equal to 1. If VIP is over 1, it means the spectra make significant influence on Y-predicted-value decision in the model.

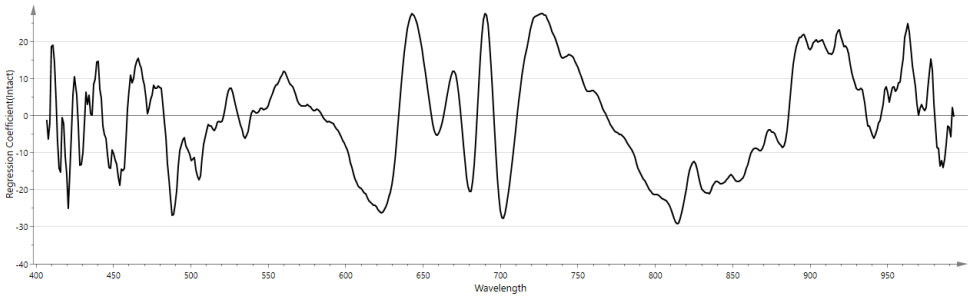

Figure 16. Regression Coefficient: Coefficients of the regression equations formulated with components based on PLS, OPLS-DA. Regression line over Y=0 means its wavelengths helped seeds to be sorted sound ones and vice versa. If lines shape sharp hills, it is meant to form a certain material's absorption peak.

제 5 장 고 찰

제 1 절 주성분분석 상에서 이상치 처리

400~1000nm, 780~1000nm 구간대에 대한 PCA-Score plot을 기반으로 Hoteling's T^2 통계량의 95% 신뢰구간에 들지 못하는 값들을 이상치로 지정하고 제외처리 한 뒤, 이상치를 처리하지 않은 경우와 비교하였다 (Figure 17. upper).

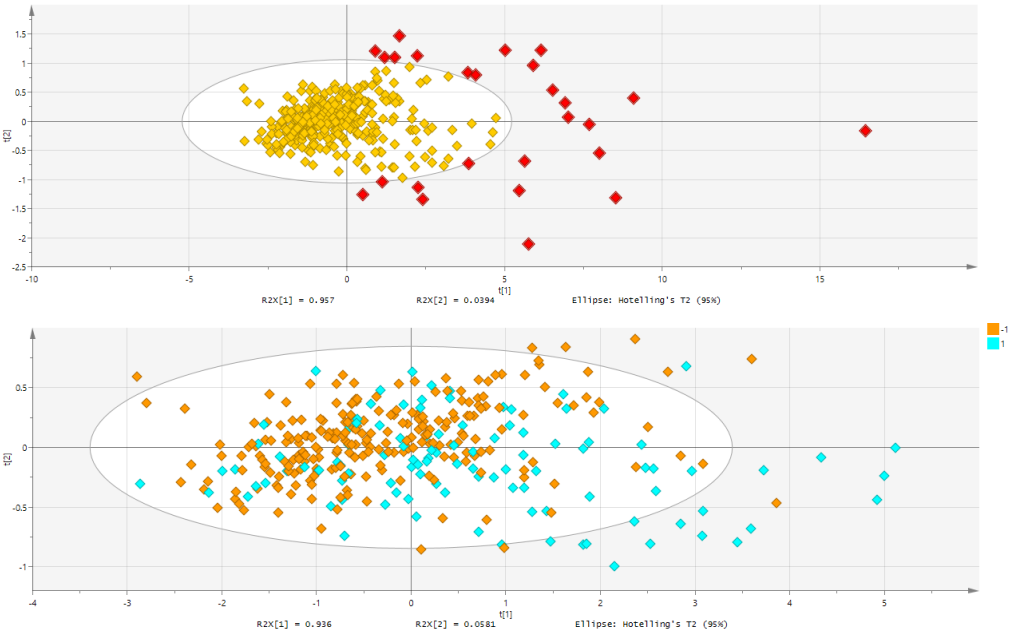

Figure 17. Schematic of outlier detection, outlier removal in 400~1000nm PCA with Hoteling's T^2 (confidence interval=95%) (upper)- Red dots are the outliers, PLS-DA analysis on raw 400~1000nm spectra after outlier detection (lower)

건전종자군의 군집 정도는 그대로였고 충해종자군 간의 거리가 가까워지

며 군집을 형성하였지만, 건전종자군과 겹치는 부분이 더 넓게 형성되었다(Figure 17. lower). 이후 전처리 없이 PLS-DA 및 OPLS-DA를 적용하였는데 400~1000nm의 경우, calibration set의 366개의 관찰값 중 26개, 780~1000nm은 18개가 이상치로 지정되어 처리되었다. 그 결과 전처리 없이 PLS-DA, 및 OPLS-DA를 적용한 결과, 정확도는 각각 86.41%, 80.98%로 기존과 동일하거나 오히려 정확도가 더 떨어지는 양상을 보였다.

적외선파장에 대한 다변량 분석 시 이상치 제거는 모델의 성능에 영향을 미치는 이상치가 존재했을 때 제거하는 것이 가장 이상적이라 판단하여(Nicolai et al., 2007), 이상치를 제거하지 않고 데이터 분석을 진행하였다.

제 2 절 파장변수대에 대한 고려

파장변수대의 사용에 있어서는 어떤 구간을 사용하는가에 따라 고려해야 할 요소가 달라진다고 보고 있다. 예컨대 정확도, F1 score, G mean 등 분류성능의 경우 780~1000nm 구간대의 SG전처리, PLS-DA를 사용한 모델이 가장 높았다. 그러나 자세한 수치로 들어갔을 경우 400~1000nm, 780~1000nm 구간대에 따른 정확도의 경우, 0.84~0.86 정도로 약 1~2% 정도의 차이를 보였으며 이는 가시광선-근적외선, 근적외선 영역대의 사용에 따른 모델 성능의 차이는 별반 없는 것을 시사한다(Figure 12).

가시광선 파장영역인 400~780nm만을 사용하였을 경우, 정확도는 66~68%로 낮았다. 불균형데이터를 고려하였을 때 민감도는 98~99%로 매우 높지만 특이도가 반영된 G mean의 경우 22~33%임을 고려하였을 때, 가시광선 파장대만을 이용하였을 경우 충해종자들까지 대부분 건전종자로 오분류가 됨을 알 수 있다. 이는 2007년 Tigabu의 연구에서 가시광선 영역의 포함여부에 상관없이 2% 오차 이내의 약 90% 정확도를 보인 것과 일치하나, 400~700nm 파장대만의 데이터를 사용하였을 때에도 약 86%의 정확도

를 보이는 것과는 대조되는 바이다(Tigabu et al., 2007).

이러한 결과가 나온 이유는 다음과 같다고 본다. 우선 가시광선영역의 경우 수많은 간섭요소가 존재하였다. 가시광선 영역 내 680nm의 경우, 가시광선대 영역 중에서도 붉은 색에 관련되어 있는 파장대의 영역이며, 엽록소의 빛 흡수로 인하여 발생하는 흡광지점이 존재한다(Wati et al., 2021). 도토리의 덜 분해된 엽록체과 도토리의 지방질 요소 등등은 가시광선 영역대의 색깔변화에 영향을 미쳤다(Bonner, 1987). 또한 충해가 심한 경우 표징으로 인하여 도토리 껍질 표면이 검게 갈변하였다. 이러한 여러 요소가 겹쳐 도토리의 경우 가시광선 영역대에서의 분석만으로는 분류를 하기 어렵게 만들었다고 판단하였다.

동시에 400~780nm의 파장변수들이 근적외선 영역대와 같이 사용되었을 경우 분류모델의 성능에 영향을 주는 것은 사실로 보인다. VIP 경우에 680nm에서의 흡광지점이 매우 큰 계수를 가지는 것을 볼 수 있으며(Figure 15), Regression Graph(Figure 16)에서도 동일한 구간에서 매우 날카로운 언덕을 그리며 흡광지점을 나타내는 것을 확인할 수 있었다. 또한 PLS의 첫 번째 주성분(77%)의 경우 600nm 이후의 가시광선, 근적외선 파장변수들로 구성된 것을 볼 수 있다(Figure 18). 이는 차후 도토리 충해여부에 대한 추가연구 시 가시광선 영역을 포함시켜야 함을 시사한다.

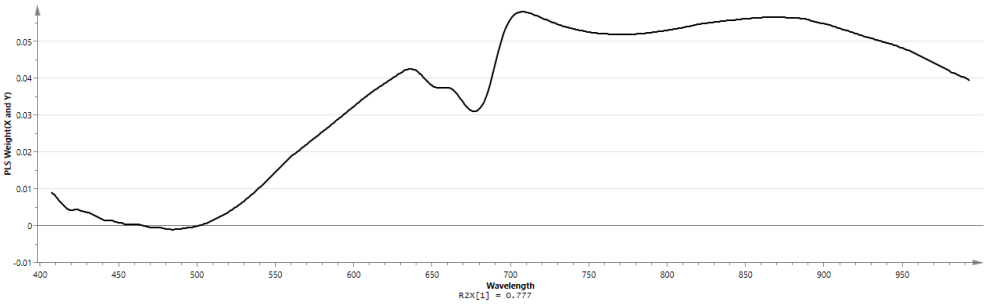

Figure 18. PLS weight of 400~1000nm Savitsky-Golay filtered PLS-DA(SGWP) model's 1st PC

근적외선대의 경우, VIP=1을 넘는 파장변수들은 대부분 960nm ~ 1000nm에 위치하였으며(Figure 15), Regression graph의 수치가 높은 흡광지점들 또한 같은 영역에 위치하였다(Figure 16). 960nm ~ 1000nm에서 발생하는 흡광지점은 분자의 O-H기 및 C-H의 수축운동으로 발생하는 3,4 번째 배음대(overtone)의 영향으로 인해 발생하는 곳으로 알려져 있으며 (Cen and He, 2007) 특히 과실류 등에서는 물과 탄수화물의 O-H기들이 이에 반응한다고 알려져 있다(Guo et al., 2016).

도토리의 주요 구성 성분은 약 7.5%의 타닌, 10%의 기타성분, 11.1%의 수분과 72%의 탄수화물(특히 전분질)로 탄수화물, 수분함량이 많은데 (정동효 등, 1975), 해충이 내부를 가해하는 과정의 영향을 받은 것으로 보인다 (Wang et al., 2010; Jamshidi et al., 2019; Jamshidi, 2020). 한편 도토리의 탄닌 성분이 가지는 흡광지점인 938nm의 경우(Takahashi et al., 2011), regression graph에서 확인이 되었으나 충해 여부에 대한 유의미한 영향은 주지 못했다.

제 3 절 전처리 및 다변량 분석기법의 처리 효과

전처리의 경우, G mean에 따른 차이가 두드려졌는데, 이는 G mean을 구성하는 특이도에 의한 것으로 보였다. 780~1000nm 구간대의 경우 전처리별 특이도의 차이, 즉 충해종자를 제대로 분별하는 정도는 62~67%로 크게 차이가 나지 않았다. 그러나 400~1000nm 구간대의 경우, SNV 및 MSC를 사용하였을 시, PLS, OPLS-DA 모두 1.64%로 충해종자를 거의 분별할 수 없었다(Figure 19).

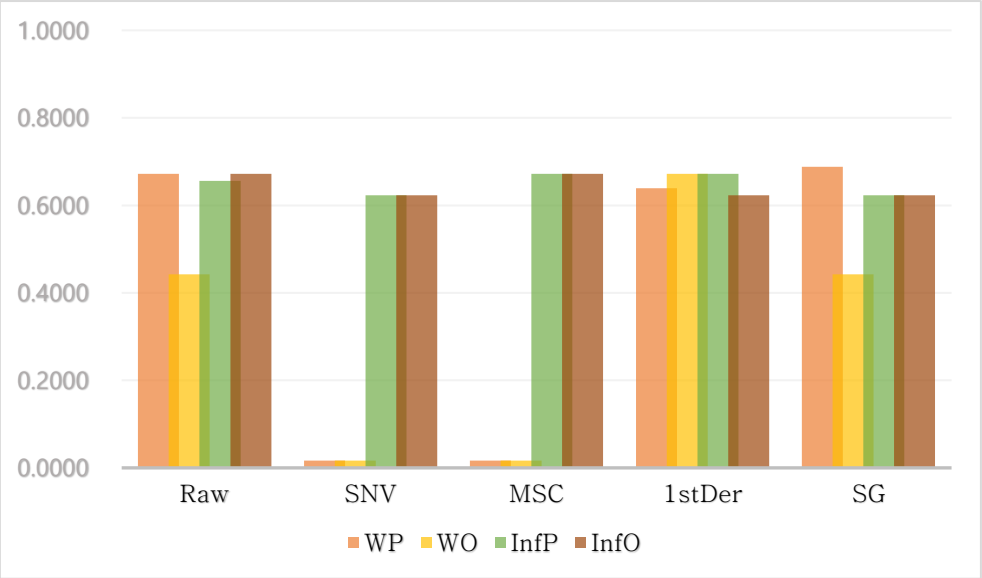

Figure 19. Specificity by preprocessing methods, W: 400~1000nm Inf: 780~1000nm P: PLS-DA O: OPLS-DA

본 결과는 SNV, MSC 전처리 방법의 특성으로 인하여 발생하였다고 판단하였다. SNV, MSV는 수식상 사용하는 변수는 다르지만 파장변수의 분산 중 산란효과(Scatter effect)로 인하여 발생하는 불필요한 잡음을 제거해주는 효과는 동일하기 때문에 실제 분석에 사용시 동일한 통계적 전처리 효과를 가지고 있다(Rinnan et al., 2009). 이러한 전처리 효과는 분산 내 산란효과의 비중이 크지 않고 화학적구성으로 인한 분산효과가 더 클 경우 구성물질 및 원소의 차이에 의한 분산 효과까지 제거할 수 있다(Næs et al., 2002). 그런데 앞서 파장변수대에 관한 고찰에서 언급하였듯이, 400~780nm 구간대의 가시광선 영역대는 도토리의 색상 변화 및 충해의 표징 등 수많은 요소에 의해 간섭을 받고, 780~1000nm 구간대의 경우 수분 및 탄수화물을 구성하는 분자에 의하여 영향을 받았다.

SNV, MSC에서 사용하는 통계량들인 평균, 표준편차(수식{3}), 상가적 및 상호작용(수식{4}) 등은 가시광선 영역상의 색상 변이뿐만 아니라 근적외선

영역에서의 구성물질의 변이 또한 포함하고 있기 때문에, 결국 최종적으로는 분류에 사용하고자 하는 데이터들까지 산란효과와 함께 제거된 듯하였다.

반면 1차 미분처리, SG 등 다른 전처리들의 경우 기준선효과를 제거하고 흡광지점을 증폭하되 산란효과를 포함한 광범위한 형태의 분산을 제거하지는 않았기 때문에 가시광선, 근적외선 영역 둘 다 특이도가 비교적 높게 나온 것으로 보인다.

결국 도토리를 대상으로 분광분석 시, 가시광선 영역까지 포함해서 품질분석에 있어서 SNV 및 MSC를 이용한 통계적 전처리는 적합치 않은 것으로 사료된다.

제 4 절 연구의 의의 및 한계

본 결과들은 도토리의 품질을 결정하는 중요한 요인들 중 하나인 충해여부를 검정하는 데 기여할 수 있을 것이다. 본 연구를 통해 확인한 유의미한 전처리 조합 및 파장대는 차후 Visible-Shortwave Infrared(400~2500nm) 구간을 사용하여 난대상록 활엽수종의 도토리를 연구하는 데 있어 참고할 만한 자료가 될 수 있을 것이다. 또한 본 연구에서 파악한 유의미한 흡광지점 및 파장변수들은 초분광, 다차원 분광분석 등의 같은 파장대 기반의 다른 광학적 분류 방법을 연구할 때 사용할 파장대를 좁히는 데 도움이 될 수 있을 것이다. 초분광, 다차원 분광분석의 경우 종자정선의 자동화에 있어 실용성이 있는 장치이지만(Feng et al., 2019), 데이터 분석 방법의 복잡성 때문에 주요한 파장변수의 파악이 매우 중요하게 사용된다.

한편, 본 연구에서는 한계를 가지고 있으며 차후 개선된 필요를 가지고 있다. 특히 충해에 따른 분류에 대해 논의가 필요하다고 생각한다. 일부 연구에서는 종자의 충해 여부가 심각하지 않을 경우 발아율에는 영향을 미치지 않았다(Guo et al., 2009). 그러나 과거의 연구는 발아율, 활력 등에 연관이

되어 있다 판단하여 충해 여부에 따라 종자 정선을 하였으며(Bonner, 1987), 실제로 cotyledon 등의 종자 내 필수 기관이 가해를 입은 경우 종자의 발아율이 감소하였다(Hou et al., 2010). 또한 지구온난화로 인해 가뭄이 예상되는 환경에서는 심각한 충해 피해 뿐만 아니라 경미한 충해 피해도 도토리 활력의 감소에 영향을 줄 것으로 예측되고 있다(Bartlow et al., 2018). 그렇기에 추후 연구를 할 시, 외관상 충해 여부에 따른 분류를 넘어서서 cotyledon의 피해 여부와 피해도 정도에 따른 사전 분류를 할 필요가 있어 보인다.

제 5 절 후속 연구에 대한 제언

파장대 및 전처리를 사용한 모델에 대해서는 전체적인 정확도나 충해종자에 대한 검정의 정확도, 즉 특이도를 높일 수 있는 개선의 여지가 있다고 보고 있다. 우선 분광분석 이외의 다른 형태의 자료를 다변량 분석 시 같이 적용하는 방법이 있을 것이다. Bianchini는 *Jatropha curcas*의 열매인 대추에 대하여 종자 내부의 여러 원인으로 인하여 비어 있는 공간이나 종자 가해 정도를 X-ray 검사를 통해 세 class로 분류하고 해당 분류와 분광분석 데이터를 동시에 활용하여 98%의 정확도로 대추의 활력 여부를 검정하였다(Bianchini et al., 2021). 도토리도 마찬가지로 X-ray 검사를 통하여 충해종자를 질적으로 분류할 수 있으며(Lombardo and McCarthy, 2009), 이러한 X-ray 데이터와 분광분석 데이터를 동시에 활용하여 도토리 충해 검정에 대한 정확도 및 특이도 등을 높일 수 있을 것이라고 기대하고 있다.

두번째로는 실험상에서의 분류군을 보다 정확하게 나누는 방법이다. Jamshidi는 충해를 입은 석류에 대하여 건전(healthy), 충해군A(unhealthy-A, 겉으로 충해 피해 발견 불가능), 충해군B(unhealthy-B, 충해피해 발견)로 분류군을 나눈 뒤 PCA-DA 모델을 적용한 사전연구가 있었으며, 모든 분류군의 예측에 대하여 92%의 정확도를 보였다(Jamshidi et al., 2019). 도토

리도 이와 마찬가지로 충해종자 분류내에서도 가해를 당하는 과정에서 표피가 부패한 경우와 탈출공만 존재하고 부패하지 않은 종자들이 있었다. 그리고 이러한 종자들은 분류 검정 모델 상 주로 사용된 660-730nm와 960-1000nm에서도 현저히 다른 양상을 보여주었다. 추후 연구 시, 충해종자를 좀 더 자잘한 분류군으로 세분화하여 검정 모델의 정확도를 높일 수 있을 것으로 보인다.

마지막으로 데이터 분류군의 불균형성을 해소하는 방법이 있다. 내부가 부패한 사과에 대하여 건전, 충해군으로 분류군을 나눈 이후, SMOTE(Synthetic minority oversampling technology) 기법을 이용하여 충해 분류군의 크기를 건전군과 같은 크기까지 증가시켰다. 이후 유의미한 파장변수만을 선택하여 따로 적용하는 과정-CARS(Competitive adaptive reweighted sampling algorithm), SPA(Successive projections algorithm)을 거쳐서 충해 여부와 상관없는 파장변수들을 제거하는 과정을 거치고 PLS-DA, LS-SVM 등의 다변량 분석기법을 적용하여 90% 이상의 정확도와 G mean을 확보하였다(Zhang et al., 2022). 본 연구의 건전, 충해종자 분류군 비율은 약 2.5:1 이므로 데이터 calibration set의 충해종자 분류군에 대한 SMOTE 기법을 적용하였을 경우 모델의 정확도 및 특이도를 높일 수 있을 것으로 생각된다.

제 6 장 결 론

해당 연구는 근적외선 구간대에서의 분광분석을 이용하여 붉가시나무 종자의 충해 여부를 검정할 수 있을 것이라는 가설을 세우고 실험을 진행하였다. 가설 검정을 위하여 400~1000nm 구간의 가시광선-근적외선 영역(VIS-NIR) 일부 구간의 분광분석 데이터를 이용하여 붉가시나무의 종자 가해여부를 판별할 수 있는지를 알아보고 판별에 가장 적합한 모델 및 파장변수를 찾고자 하였다.

이를 위하여 400~1000nm 구간 전체를 사용할 것인지, 아니면 780~1000nm의 근적외선 영역대만 사용할 것이 좋은가, 어떠한 전처리 방법을 사용하는 것이 분광분석 데이터의 잡음 제거에 효과적이며, 어떠한 다변량 분석 방법을 사용하는 것이 가장 효과적인 판별 모델 설계에 적합한 것인지를 알아보았다. 그를 위하여 정확도 및 정밀도, F1 Score, G mean 등의 모델 평가 지표를 사용하여 모델들의 성능을 파악하였다. 마지막으로 VIP(Variable important for the projection)식을 이용하여 해당 모델에서 어떤 파장대의 값이 더 높은 영향력을 가지는 가를 판별하여 추후 유효한 파장선별 시 어떤 파장변수들을 사용해야 하는지를 조사 및 비교하였다.

모든 파장대 및 근적외선만을 사용한 모델들은 0.86 ~ 0.87의 정확도 및 0.8~0.82 근처의 G mean의 차이가 얼마 안나는 분류성능을 보여주었다. 반면 가시광선 파장만을 사용하였을 경우, 충해종자에 대한 검정력이 0에 가깝게 떨어지며 0.66의 낮은 정확도를 보여주었다. 최종적으로 400 - 1000nm 파장대에서 Savitsky - Golay 전처리를 한 뒤 PLS-DA를 이용한 모델의 정확도가 0.8641, 정밀도 0.8603, F1 Score 0.9035, G mean 0.8093으로 가장 우수한 지표를 보여주었다.

VIP Score 및 Regression를 이용하여 모델 내 파장변수의 영향력을 확인

한 결과 660-720nm, 960-1000nm의 파장변수 및 흡광지점이 분류 모델에 유의미한 영향을 미치는 것을 알 수 있었다. 가시광선 영역대의 경우 충해 여부 이외의 많은 요소로 인하여 간섭이 발생하나 모델의 검정력에 큰 영향력을 주었다. 근적외선 영역대의 경우 유충이 종자를 가해하는 과정에서 발생하는 수분 및 탄수화물의 변화의 차이가 모델의 분별에 영향을 미친 것으로 분석되었다. 결국 파장대의 경우, 400~1000nm 구간대 전체를 사용한 것이 적절해 보인다. 또한 추후 파장변수의 개수를 줄이고자 할 때 앞서 언급한 두 흡광지점을 중심으로 좁혀 나가는 것이 중요해 보인다.

전처리의 경우, G mean 및 G mean을 구성하는 특이도에 따라 차이가 발생하였다. 근적외선 구간대의 경우 65~67%로 차이가 적었으나, 400~1000nm을 파장구간대에서 SNV 및 MSC 전처리의 경우 충해종자 분별능력이 1.64%로 매우 낮았다. 이는 SNV, MSC 전처리를 이용하여 산란효과를 제거하는 과정에서 충해 분류에 필요한 수분 및 탄수화물로 인해 발생하는 분산까지 제거된 것으로 분석된다. 결국 가시광선 영역까지 포함하여 분류모델을 만들 경우 산란효과를 제거하는 과정을 거치는 전처리 방법은 주의할 필요가 있어 보인다.

본 연구의 전처리 및 파장대조합은 1000nm 이상의 2500nm까지의 더 넓은 파장대의 연구 시 참고할 수 있을 것이며, 연구를 통해 얻은 유의미한 파장대 및 흡광지점은 다차원 분광분석 등의 기기를 이용하는 데 있어 연구범위를 줄여줄 수 있을 것이다. 한편 해당 연구를 개선하기 위해서는 X-ray 등의 다른 분류자료를 같이 이용하여 분석하거나 충해분류군의 세분류화를 통해 도토리에 대한 충해 검정 모델의 정확도를 개선할 수 있을 것으로 보고 있다.

Abstract

Along with the global carbon neutral trend, the need for carbon absorption sources, especially broad-leaved trees that can adapt to climate change, is increasing. *Quercus acuta* Thunb is an evergreen broad-leaved tree, and it is expected that its potential distribution is expanding from the coast to the inland of Korea in the future dew to climate change. To efficiently produce seeds for future afforestation demands, low-cost, non-destructive seed testing using spectroscopy should be investigated. However, studies on the quality classification of evergreen broad-leaved seeds are insufficient, and research on forest seeds using near-infrared spectroscopy is lacking in Korea. Therefore, this study was conducted to distinguish between sound and damaged acorns using vis-near-infrared spectroscopy (400~1,000nm) based on preliminary research on chestnuts and pomegranates. A total of 550 acorns were collected from a *Quercus acuta* seed orchard located in Seogwipo-si, Jeju-do. The collected seeds were classified into 362 sound seeds and 178 damaged seeds through visual investigation. The seed lots were then randomly divided into a calibration set of 366 acorns (246 sound and 120 damaged) and a validation set of 184 acorns(116 sound and 58 damaged), respectively. Using Ocean-optics' Flame-S-Vis-Nir spectrometer and ISP-REF, 550 spectra were obtained by averaging 10 repeated observations at each cardinal point on the seeds, with an average of 550 observations. Afterwards, data preprocessing methods such as first derivative, MSC, SNV, and Savitsky-Golay filter were used to reduce the error of the

spectroscopic data. In addition, PLS-DA and OPLS-DA multivariate analysis methods were applied. The performance of models was compared between different wavelength ranges, preprocessing, and multivariate analysis methods. Model evaluation indexes such as accuracy, precision and variable importance in projection were used to measure the performance. The VIS-NIR spectra of acorns displayed a sharp absorption peak at 660-720 nm and a smooth downward absorption peak at 900~1,000 nm. As a result of multivariate analysis and model evaluation, the best model to distinguish damaged seeds was analyzed with PLS-DA and SG treatment in the whole range of 400 to 1,000 nm. The model showed accuracy of 0.8641, precision of 0.8603, F1 Score of 0.9035, and G mean of 0.8093 ($p<0.05$). The absorbance peaks at 660-720 nm and 960-1,000 nm had a significant effect on the performance of the model. In the latter case, it was presumed that the acorn worms affected moisture and carbohydrates. In the former case, the peak was strongly influenced by chlorophyll and color variation. However, the performance of the model changed when the 660-720nm peak was used with a near-infrared wavelength band (780~1000nm). The results for absorption peaks can be used to narrow the scope of subsequent studies utilizing wider wavelength bands or multidimensional spectroscopy. To improve classification accuracy, this study suggested incorporating other types of data, such as x-rays, subdividing classification groups of insect-damaged seeds or oversampling the minority classes.

참고문헌

구경아, 공우석, 김종규. 2001. 한반도 상록활엽수의 지리적 분포와 기후요소, 대한지리학회지 36: 247-257.

국립산림품종관리센터 2021. 제 2차 채종원 조성 및 관리 추진계획(2022~2026). 국립산림품종관리센터, 충주, 21-31.

김대용, 조병관, 이윤수. 2012. 광 반사방식을 이용한 감염 씨감자 비파괴 선별 기술 개발, 농업과학연구 39: 117-123.

미래전략과 2020. 2050 탄소중립 추진전략. 기획재정부, 서울.

민태기, 강우식. 2003. 근적외선을 이용한 박의 단일 종자 단위로 발아 종자와 불발아 종자의 비파괴적인 선별, 원예과학기술지 44: 545-548.

박선욱, 구경아, 공우석. 2016. 기후변화에 따른 한반도 난대성 상록활엽수 잠재서식지 분포 변화, 대한지리학회지 51: 201-217.

산림청 2021. 2021년도 산림자원분야 사업계획. 산림청, 대전.

신만석, 서창완, 박선욱, 홍승범, 김진용, 전자영, 이명우. 2018. 분산 능력을 고려한 기후변화에 따른 붉가시나무의 잠재서식지 분포변화 예측연구, 환경영향평가 27: 291-306.

이기원, 송요욱, 김지혜, 라하만, 아티쿨, 오미래, 박형수. 2020. 근적외선분광법을 이용한 수수× 수단그라스 교잡종 종자의 품종 판별, 한국초지조사료학회지 40: 259-264.

이정현, 최병희. 2010. 상록활엽수 3종의 한반도분포 및 자생북한계지, 식물분류학회지 40: 267.

정동효, 유태종, 최병규. 1975. 도토리 녹말의 이용에 관한 연구-제 1 보 도토리 녹말의특성, 한국응용생명화학회지 18: 102-108.

정회일, 김효진. 2000. [해설] 근적외선 분광법의 원리, 분석과학 13: 138-151.

최광식. 2006. 참나무 구과를 가해하는 해충, 조경수: 9-11.

Ballabio, D. and Consonni, V. 2013. Classification tools in chemistry. Part 1: Linear models. PLS-DA, Analytical methods 5: 3790-3798.

Bartlow, A. W., Agosta, S. J., Curtis, R., Yi, X. and Steele, M. A. 2018. Acorn size and tolerance to seed predators: the multiple roles of acorns as food for seed predators, fruit for dispersal and fuel for growth. Integrative Zoology 13: 251-266.

Bianchini, V. d. J. M., Mascarin, G. M., Silva, L. C. A. S., Arthur, V., Carstensen, J. M., Boelt, B. and Barboza da Silva, C. 2021. Multispectral and x-ray images for characterization of *Jatropha*

curcas L. seed quality, Plant Methods 17: 9-9.

Bonner, F. T. 1987. Seed biology and technology of *Quercus*. Louisiana, U.S., US Department of Agriculture, Forest Service, Southern Forest Experiment Station. pp. 5-6, 11-12.

Cen, H. and He, Y. 2007. Theory and application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science and Technology 18: 72-83.

Connor, K. 2004. Storing acorns, Native Plants Journal 5: 160-166.

Daneshvar, A., Tigabu, M., Karimidoost, A. and Oden, P. C. 2015. Single seed near infrared spectroscopy discriminates viable and non-viable seeds of *Juniperus polycarpos*. Silva Fennica 49: 14.

Dorsey, C. K., Tryon, E. H. and Carvell, K. L. 1962. Insect damage to acorns in West Virginia and control studies using granular systemic insecticides. Journal of Economic Entomology 55: 885-888.

FAO. 2010. Seeds in emergencies: a technical handbook. Rome, Italy. FAO.

Farhadi, M., Tigabu, M., Stener, L.-G. and Odén, P. C. 2016. Feasibility of visible+ near infrared spectroscopy for non-destructive verification of European× Japanese larch hybrid seeds. New Forests 47: 271-285.

Farrés, M., Platikanov, S., Tsakovski, S. and Tauler, R. 2015. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. Journal of Chemometrics 29: 528-536.

Feng, L., Zhu, S., Liu, F., He, Y., Bao, Y. and Zhang, C. 2019. Hyperspectral imaging for seed quality and safety inspection: a review. Plant Methods 15: 91-91.

Gribko, L. S. and Jones, W. E. 1995. Test of the float method of assessing northern red oak acorn condition. Tree Planters' Notes 46: 143-147.

Guo, C.-R., Lu, J.-Q., Yang, D.-Z. and Zhao, L.-P. 2009. Impacts of burial and insect infection on germination and seedling growth of acorns of *Quercus variabilis*. Forest Ecology and Management 258: 1497-1502.

Guo, W., Zhao, F. and Dong, J. 2016. Nondestructive measurement of soluble solids content of kiwifruits using near-infrared hyperspectral imaging. Food Analytical Methods 9: 38-47.

Hirayama, D., Fujii, T., Nanami, S., Itoh, A. and Yamakura, T. 2012. Two-year cycles of synchronous acorn and leaf production in biennial-fruiting evergreen oaks of subgenus *Cyclobalanopsis* (*Quercus, Fagaceae*). Ecological Research 27: 1059-1068.

Hou, X., Yi, X., Yang, Y. and Liu, W. 2010. Acorn germination and seedling survival of *Q. variabilis*: effects of cotyledon excision, Annals of Forest Science. 67: 711-711.

Hwang, J.-S., Kang, J.-T., Son, Y.-M. and Jeon, H.-S. 2015. Prediction of the optimal growth site and estimation of carbon stocks for *Quercus acuta* in Wando Area, Journal of Climate Change Research 6: 319.

Jamshidi, B. 2020. Ability of near-infrared spectroscopy for non-destructive detection of internal insect infestation in fruits: Meta-analysis of spectral ranges and optical measurement modes, Spectrochimica acta. Part A. Molecular and Biomolecular Spectroscopy 225: 117479.

Jamshidi, B., Mohajerani, E., Farazmand, H., Mahmoudi, A. and Hemmati, A. 2019. Pattern recognition-based optical technique for non-destructive detection of *Ectomyelois ceratoniae* infestation in pomegranates during hidden activity of the larvae. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 206: 552-557.

Jeni, L. A., Cohn, J. F. and De La Torre, F. 2013. Facing imbalanced data-recommendations for the use of performance metrics. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, New York, U. S. IEEE: 245-251.

Kang, W. S. 2009. Nondestructive determination of seed viability by optical methods. Doctoral thesis. The Graduate School, Daegu University, Daegu.

Khan, S. and Habib, A. 2015. Effect of pre-sowing treatments on seed germination in *Quercus glauca* Thunb., collected from different sampling sites of the Himalayan region of Pakistan, International Journal of Biosciences 6: 42-48.

Kim, H.-J. and Lee, S.-H. 2017. Estimating carbon storage and CO2 absorption by developing allometric equations for *Quercus acuta* in South Korea. Forest Science and Technology 13: 55-60.

Kim, S. and Park, I.-H. 2021. Acorn production and characteristics of

Quercus acuta thunb-Focused on Wando, Jindo and Haenam in Jeollanam-do, Korea. Korean Journal of Environment and Ecology 35: 621-631.

Lombardo, J. A. and McCarthy, B. C. 2009. Seed germination and seedling vigor of weevil-damaged acorns of red oak. Canadian Journal of Forest Research 39: 1600-1605.

Min, T. G. and Kang, W.-S. 2008. Nondestructive classification of viable and nonviable radish (*Raphanus sativus* L.) seeds using single seed near infrared spectroscopy. Horticulture Environment and Biotechnology 49: 42-46.

Moscetti, R., Haff, R. P., Aernouts, B., Saeys, W., Monarca, D., Cecchini, M. and Massantini, R. 2013. Feasibility of Vis/NIR spectroscopy for detection of flaws in hazelnut kernels. Journal of Food Engineering 118: 1-7.

Moscetti, R., Haff, R. P., Saranwong, S., Monarca, D., Cecchini, M. and Massantini, R. 2014. Nondestructive detection of insect infested chestnuts based on NIR spectroscopy. Postharvest Biology and Technology 87: 88-94.

Nehir, M., Frank, C., Assmann, S. and Achterberg, E. P. 2019. Improving optical measurements: Non-linearity compensation of compact charge-coupled device (CCD) spectrometers. Sensors 19.

Næs, T., Isaksson, T., Fearn, T. and Davies, T. 2002. A user-friendly guide to multivariate calibration and classification. Chichester, UK, NIR Publications. pp. 114-118.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I. and Lammertyn, J. 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biology and Technology 46: 99-118.

Novikov, A. I. 2019. Visible wave spectrometric features of scots pine seeds: the basis for designing a rapid analyzer. 2020 IOP Conference Series: Earth and Environmental Science.: Earth Environmental Science. Bristol, UK. IOP Publishing. 226: 12064.

Novikov, A. I., Ersson, B. T., Malyshev, V. V., Petrishchev, E. P. and Ilunina, A. A. 2020. Mechanization of coniferous seeds grading in Russia: a selected literature analysis. 2020 IOP Conference Series: Earth and Environmental Science.: Earth Environmental Science. Bristol. IOP Publishing. 595: 12060.

Novikov, A. I. and Novikova, T. P. 2018. Non-destructive quality control of forest seeds in globalization: Problems and prospects of output innovative products. Globalization and Its Socio-Economic Consequences: 1260-1267.

Rinnan, Å., Van Den Berg, F. and Engelsen, S. B. 2009. Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends in Analytical Chemistry 28: 1201-1222.

Shenk, J. S. and Westerhaus, M. O. 1993. Near infrared reflectance analysis with single and multiproduct calibrations. Crop Science 33: 582-584.

Son, Y. M., Kim, R. H., Kang, J. T., Lee, K. S. and Kim, S. W. 2014. A practical application and development of carbon emission factors for 4 major species of warm temperate forest in Korea. Journal of Korean Forest Society 103: 593-598.

Takahashi, A., Shimada, T. and Kawano, S. 2011. Nondestructive determination of tannin content in intact individual acorns by near-infrared spectroscopy. Ecological Research 26: 679-685.

Tigabu, M., Daneshvar, A., Wu, P., Ma, X. and Christer Odén, P. 2019. Rapid and non-destructive evaluation of seed quality of Chinese fir by near infrared spectroscopy and multivariate discriminant analysis. New Forests 51: 395-408.

Tigabu, M., Fjellström, J., Odén, P. C. and Teketay, D. 2007. Germination of *Juniperus procera* seeds in response to stratification and smoke treatments, and detection of insect-damaged seeds with VIS + NIR spectroscopy. New Forests 33: 155-169.

Tigabu, M. and Odén, P. C. 2002. Multivariate classification of sound and insect-infested seeds of a tropical multipurpose tree, *Cordia africana*, with near infrared reflectance spectroscopy. Journal of Near Infrared Spectroscopy 10: 45-51.

Trygg, J. and Wold, S. 2002. Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics: A Journal of the Chemometrics Society 16: 119-128.

Wang, J., Nakano, K., Ohashi, S., Takizawa, K. and He, J. 2010. Comparison of different modes of visible and near-infrared spectroscopy for detecting internal insect infestation in jujubes. Journal of Food Engineering 101: 78-84.

Wati, R. K., Pahlawan, M. F. R. and Masithoh, R. E. 2021. Development

of calibration model for pH content of intact tomatoes using a low-cost Vis/NIR spectroscopy. 2021 IOP Conference Series: Earth and Environmental Science. Earth Environmental Science. Bristol. IOP Publishing. 686: 12049.

Wise, B. M., Gallagher, N. B., Bro, R., Shaver, J., Windig, W. and Koch, R. S. 2006. PLS Toolbox 4.0. Wenatchee, U.S., Eigenvector Research, Inc. pp. 157-160.

Zhang, Z., Liu, H., Chen, D., Zhang, J., Li, H., Shen, M., Pu, Y., Zhang, Z., Zhao, J. and Hu, J. 2022. SMOTE-based method for balanced spectral nondestructive detection of moldy apple core. Food Control 141: 109100.