농학석사 학위논문

식물표본 DB 오류 현황과 원인 분석

-국립수목원과 국립생물자원관 소장 일부 목본 분류군을 중심으로-

Status assessment and cause of herbarium database errors

-Selected woody plants taxa stored in national herbarium of Korea-

2017 년 8 월

서울대학교 대학원
산림과학부 산림환경학 전공
김 혜 원

식물표본 DB 오류 현황과 원인 분석

– 국립수목원과 국립생물자원관 소장
일부 목본 분류군을 중심으로 –

Status assessment and cause of
herbarium database errors

–Selected woody plants taxa
stored in national herbarium of Korea–

지도교수 장 진 성

이 논문을 농학석사 학위논문으로 제출함
2017 년 7 월

서울대학교 대학원
산림과학부 산림환경학 전공
김 혜 원

김혜원의 농학석사 학위논문을 인준함
2017 년 7 월

위 원 장 강 주석 (인)
부위원장 [illegible] (인)
위 원 김 희 (인)

초 록

표본은 수 백 년 간의 자연사 정보를 축척하고 있으며 생물다양성 연구에 있어 핵심적인 정보를 제공해 왔고 현재는 표본 정보를 디지털화 하여 온라인에서 제공함으로써 그 가치가 제고되고 있다. 국내에서도 2000년대 전후를 시작으로 표본정보의 디지털화와 데이터베이스 관련 사업이 추진되고 있다. 그러나 시스템의 구축이나 자료의 양을 늘리는데 비해 자료의 질에 대한 관심, 즉 자료 정제의 중요성에 대한 인식은 부족하다. 이에 본 연구에서는 국가기관 표본DB와 표본의 관리 현황에 대한 정보를 제공하고, 오류를 점검하며, 오류의 유형 분석을 통해 원인을 파악해 보며 부분적으로나마 오류를 정제할 수 있는 방법을 제시하고자 하였다. 대상은 국립수목원 산림생물표본관(KH)과 국립생물자원관(KB)에 소장된 식물표본과 표본정보로 양 기관에서 목본 17,517 점의 좌표정보를 포함한 식물표본정보(일차종발생정보)를 제공받아 DB형식으로 전환하고 직접 표본관을 방문하여 표본을 동정하였다. 공간적 오류의 경우 좌표 정보의 부재가 문제가 되었다. 분류학적인 오류의 경우 조사대상 전체의 평균 오동정률은 10.4%정도였는데, 오동정이 없는 분류군부터 67.07%로 매우 높은 분류군들이 섞여 있으며 오동정률에 있어 분류군별 경향성을 찾기가 어려웠다. 이는 자료를 사용함에 있어 신뢰도를 크게 떨어뜨리며 자료 사용 자체에 대한 혼란을 야기할 수 있다. 오류의 정제에 있어 공간적인 오류는 지리참조연산을 이용해 정제가 가능하며 분류학적인 오류는 분포지역 비교 및 분포도 작성을 통해 25.6% 정도

를 탐지할 수 있었다. DB의 관리 현황 점검에서 국립수목원 물푸레나무속과 수수꽃다리속 DB와 실제 표본을 대조, 비교해본 결과 DB와 실제 표본 사이에 30-38% 정도의 차이가 있었다. DB관리에 있어 핵심은 자료의 데이터베이스화로 표본의 채집부터 디지털화, 오류의 점검, 정제와 예방의 모든 과정은 데이터베이스를 통해 이루어져야 한다.

주요어 : 식물표본정보, 일차종발생정보, 표본관 DB, 자료 품질, 자료 정제, 공간적 오류, 분류학적 오류, 오류의 탐지

학번 : 2015-23017

목 차

List of Table

List of Figure

제 1 장 서론

1. 연구 배경

표본관은 수백 년 간의 자연사자료를 축적해 둔 공간으로서 생물다양성 연구에 있어 핵심적인 정보를 제공해 왔다. 표본의 정보는 해부학적, 형태학적, 유전적, 지리학적 정보 등을 통합적으로 제공하므로 종에 대한 전반적인 이해를 가능하게 해준다. 또한 관측 정보와는 달리 표본

이라는 실체가 있으므로 DNA분석과 같은 현대적인 기술을 이용하여 몇 번이고 실험이나 조사를 할 수 있다(Scoble and Bourgoin 2010). 실용적인 측면에 있어서는 어떤 연구할 경우 새롭게 채집을 해서 자료를 수집하는 것에 비해 기존의 표본 자료를 이용할 경우 비용과 시간을 크게 줄여준다(Fuentes et al., 2013). 또한 지리적으로 멀리 떨어진 지역의 종이라든가 접근하기 어려운 종, 멸종한 종, 또는 이제는 다시 돌리거나 복원할 수 없는 환경에 존재했던 종의 연구에 있어서 표본만의 귀중한 가치가 특히 빛을 발한다(Guerin 2013).

현재 우리나라 국립표본관은 두 곳이 있는데 산림청 소속의 국립수목원 산림생물표본관(KH)은 2003년 설립된 전용표본관으로 51만여점의 식물표본을 비롯하여 곤충 45만여 점, 버섯과 지의류 및 기타 4만여 점이 소장되어 있다. 환경부 소속의 국립생물자원관 생물표본 수장고

(KB)는 2007년 개관한 국립생물자원관 내에 위치하고 있으며 관속식물 60만여 점, 곤충 66만여 점, 동물 110만여 점, 기타 35만여 점이 소장되어 있다.

이러한 자료의 정보를 어떻게 사람들이 쉽게 접근하게 하고 이용할 수 있게 하느냐는 것이 문제인데 이에 표본 정보를 디지털화하여 데이터베이스로 만들어 제공하는 작업이 답으로 제시되고 있다(Chavan and Krishnan 2003). 영국의 Kew garden은 Kew Herbarium Catalogue(http://apps.kew.org/herbcat/gotoHomePage.do), 호주의 AVH(The Australasian Virtual Herbarium, http://avh.chah.org.au/), 미국의 Harvard University Herbaria(https://huh.harvard.edu/) 등 세계의 주요 표본관과 박물관들은 웹상에서 온라인 표본관을 운영하며 모든 정보를 제공하고 있다. 전세계 생물종발생자료를 제공하고 있는 GIBF(Global Biodiversity Information Facility, http://www.gbif.org)는 생물다양성데이터를 전세계 누구나 이용할 수 있도록 하는 것을 목적으로 1999년에 OECD의 인준을 받아 2001년에 설립되었으며 2017년 6월 현재 전세계 783백만 개의 종발생정보를 제공하고 있다.

국내에서도 2000년대 전후로 표본관의 표본 정보 즉, 일차종발생정보의 디지털화가 꾸준히 진행되어 왔다. 국립수목원은 국가생물종지식정보시스템(http://www.nature.go.kr)을 운영하고 있으며 국립생물자원관에서는 한반도의 생물다양성 포털(https://species.nibr.go.kr)을 구축하여 운영하는 등 국가차원에서도 표본정보의 디지털화와 데이터베이스 관련 사업을 추진하고 있다. 또한 미래창조부와 한국정보화진흥원의 국가

DB사업으로 한국생물다양성정보기구(KBIF)에서 운영하는 나리스라는 국가자연사연구종합정보시스템(NARIS, Korean Natural hIstory Research Information System)도 구축되어 있다.

이미 외국에서는 생물다양성 정보학(Biodiversity informatics)의 발전과 더불어 디지털 정보를 단순히 일차종발생정보의 제공으로만 쓰는 것이 아니라 다양한 분야에서 활용을 하고 있다. Chapman(2005)은 GBIF프로그램의 일환으로 일차종발생정보가 어떻게 활용되는지를 크게 22개의 분야별로 정리하고 설명한 바 있으며, 표본관의 72가지 용도에 대한 정리를 통해 표본정보의 유용함을 보여주는 연구도 있었다(Funk, 2003). 과거에 수집되었으나 미기록종이 아직 많기 때문에 표본관이 새로운 종을 찾아내는 개척자 역할을 한다는 연구도 있으며(Bebber et al., 2010), 지표식물의 표본을 통해 광맥을 찾아내기도 하고(Brooks et al., 1977), 기후 변화와 관련된 연구(Davis et al., 2015; Gallagher et al., 2009; Guerin et al., 2012; Hart et al., 2014; MacGillivray et al., 2010; Miller-Rushing et al., 2006), 생물지리학(Peterson et al., 1998; Schulman et al., 2007), 멸종위기종 관리 및 보전 계획의 수립(Crisp et al., 2001; Faith et al., 2001; Kier and Barthlott 2001), 보호대상지의 선정(Margules and Pressey, 2000; Williams et al., 2002), 식물의 생장 연구(McGraw, 2001), 병해충 연구(Booth et al., 2000), 농림어업생산 등의 다양한 분야에서 활용이 되고 있다. 특히 종다양성 연구에 일차종 발생정보를 적극 활용하고 있으며(Anderson, 2012; Faith et al., 2001; Funk and Richardson, 2002; Funk et al., 1999;

Graham et al., 2004; Jiménez-Valverde et al., 2010) 외래식물 관리와 대책수립을 위한 연구에서 많이 이용되고 있다(Aikio et al., 2010; Crawford and Hoagland, 2009; Delisle et al., 2003; Fuentes et al., 2013). 지구 온난화에 따른 이른 개화를 보여주는 연구(Primack et al. 2004; Panchen et al. 2012)나 이른 개화기와 육질과 유형의 관계에 대한 연구(Bolmgren and Lonnberg, 2005)도 표본을 통해 진행되었다.

일차종발생정보 자체를 연구한 경우들도 있는데 Lavoie(2013)는 전산화된 표본관이 어떤 영향을 미치는지, 즉 생물지리학과 환경연구에 있어 얼마만큼 이용되는지, 어떤 분야에서 표본의 활용이 많은지, 표본관의 규모와 활용 간의 관계가 있는지 등에 대하여 연구를 하였다. 특정 지역과 종에 대해서이긴 하지만 표본관 일차종발생정보가 얼만큼 현재의 집단 분포를 예측하게 해 주는지에 대한 연구도 진행되었다(Applequist et al., 2007). Crawford et al.(2009)은 표본 정보와 현장조사를 비교하는 연구를 통해 일차종발생정보가 실제 종분포와 종출현도를 정확하게 보여준다고 하였다..

그러나 일차종발생정보의 활용도가 높아질수록 일차종발생정보를 이용하는데 있어 여러 가지 한계점도 지적되고 있다. 표본정보와 같이 사용자가 자료를 생성하는 사람이나 과정과 동떨어져 있을 경우 자료의 품질은 대량의 자료를 이용하는 사람들에게 있어 큰 문젯거리가 될 수 있다(Goodchild and Clarke, 2002). Anderson(2012)등은 직접적으로 일차종발생정보를 사용한 많은 연구들이 의도한 종을 대표한다고 할 수 있을지 의심해봐야 한다고 하였다. 분류학적인 오동정 문제 또는 다양한

품질의 동정 정도, 부재하거나 부족한 라벨정보와 공간정보의 오류, 공간적 또는 환경적인 채집지의 편향, 채집 시기의 편향, 채집된 종의 불균형 등이 일차종발생정보를 활용하는데 걸림돌로 지적되고 있다(Anderson, 2012; Funk and Richardson, 2002; Graham et al., 2004; Hortal et al., 2007). 또한 알려진 많은 종들에 대한 적절한 분포 데이터가 부족하며(Whittaker et al., 2005) 존재 데이터와 부재 데이터의 사용에 여부에 따라 모델링의 결과가 달라지는 등에 대한 문제점을 가지고 있다(Graham et al., 2004).

사람들은 디지털화된 표본 라벨 정보를 이용할 때 그 라벨 정보가 정확할 것이라는 가정을 한다(Goodwin et al., 2015). 이렇게 무비판적으로 데이터를 사용하는 경향은 오류가 많은 결과를 낳거나 잘못된 정보를 제공하고, 틀린 연구결과를 도출할 수도 있기 때문에(Chapman, 2005) 일차종발생정보 이용의 문제점 중 자료의 오류에 대한 문제가 대두되고 있다.

Jacobs et al.(2017)은 Melaleuca종의 오류 연구에서 72%의 표본의 이름이 정확하지 않으며 이 중 30%는 이명과 관련된 명명법적인 오류였고, 70%는 오동정 등의 분류학적인 오류에 의한 것이라 하였다. 폴란드 민속식물학 표본의 신뢰도에 대한 연구에서는 약 10%의 표본이 오동정된 것으로 밝혀졌다(Łuczaj, 2010). 분류학에 있어 상대적으로 미개척의 세계로 남아있는 열대식물 표본의 경우에도 50% 이상이 부정확하게 이름 붙여졌는데 이러한 결과는 무비판적인 표본 데이터의 사용에 대한 심각성을 함축적으로 보여준다(Goodwin et al., 2015). 국내 쇠무

릎과 털쇠무릎의 연구에서 시도한 재동정의 경우, 쇠무릎이 41%, 털쇠무릎의 경우 28%가 오동정으로 확인되기도 했다(Chang et al., 2012). 종분포모델링(SDM, Species Distribution Modeling)에 있어 오동정이 관심종의 예상 분포지를 변화시킬 수 있다는 영향에 대한 우려도 있었다(Costa et al., 2015).

Soberon et al.(2004)은 현재 표본관들의 일부 좌표만이 제대로 입력이 되어 있으며 잘못된 좌표정보는 표본정보 사용에 큰 한계가 될 수 있다고 지적하였고 Tobler et al.(2007)은 자료의 질은 라벨의 내용에 크게 의존하는데 어떤 라벨은 한정된 정보만을 담고 있고 어떤 라벨은 잘못된 좌표정보를 가지고 있기도 하므로 분포에 자료를 이용할 때는 시간이 걸려도 반드시 공간 정보에 대한 확인이 필요하다고 하였다.

자료를 데이터베이스로 입력하는 비용도 상당하지만(Armstrong 1992) 잘못된 자료를 바로잡는 작업은 더 많은 비용이 들기 때문에 오류를 수정하기 보다는 오류를 사전에 예방하는 것이 비용적으로도 더 효율적이다(Redman, 2001). 오류의 예방은 오류의 탐지나 정제보다 비용이나 시간 면에서 훨씬 우월하나 오류는 발생할 수 밖에 없는 것이므로 오류의 예방과 정제 모두 데이터 관리 방침에서 중요하게 다루어져야 한다(Chapman, 2005). 오류의 예방을 위해 우선되어야 하는 것이 바로 오류의 점검이며 오류의 점검 및 정제를 통해 오류의 예방 또한 가능하다.

2. 연구 목적

현재 전세계적으로 표본관은 그 필요성에 대한 논의와 더불어 재정적인 어려움을 겪고 있으며 표본관의 정보를 적극적으로 활용하는 것이 미래의 안정적 운영을 위한 가장 좋은 방법이라 하였다(Wen et al., 2015). 단순히 데이터의 양이 많은 것이 중요한 것이 아니라 신뢰할 수 있는 데이터베이스인지를 확인시켜주는 것이 필요하며 신뢰할 수 있는 데이터베이스를 구축해 나가는 것이 필수적이다(Scott and Hallam, 2003) 정보를 활용하기에 앞서 제공할 정보의 정확성을 확보하는 것은 그 무엇보다 중요하나(Goodchild and Clarke, 2002) 현재 국가기관 차원에서의 자료의 디지털화와 정보의 축적에 힘쓰는 정도에 비해 자료정제(data cleaning)에 대한 중요성에 대한 인식은 아직 부족하다. 국내 표본 관리에 있어서도 수집과 자료 입력의 정확성에 관한 연구, 데이터 정제 방법에 대한 선행된 연구나 활동은 전무하다. 오류를 이해하는 것은 자료의 품질 통제에 직접적인 역할을 한다고 하였다(Burrough and McDonnel, 1998).

본 연구에서는 국립표본관 자료에도 많은 오류(명명적, 분류학적, 공간적)가 있을 것이라 가정하고 1) 국립표본관 자료의 오류 현황을 오류의 유형별로 살펴보고 2) 오류 중 분류학적 오류를 분석하고 원인을 파악해 보며, 3) 분포도 작성을 통해 오류를 예측하고 정제하는 것이 가능하다는 것을 보여주며 마지막으로 4) 향후 표본관 DB관리의 개선 방향에 대한 논의를 하고자 한다.

제 2 장 연구사

자료의 수집부터 입력까지 오류를 줄이려 노력하지만 대부분의 데이터에서 오류는 일반적인 것이며(Goodchild and Clarke, 2002) 데이터 오류를 피하기 위해 극도의 노력을 다 했을 경우의 오류율이 5% 또는 그 이상이라고 한다(Redman, 1998). 이런 오류들은 자료에 있어 내재적인 것이며 피하기 어려운 것이지만 오류를 최소화시킬 필요가 있기에 오류를 찾아내고 정제함으로써 자료의 품질을 관리하고자 하는 노력들이 계속되고 있다(Maletic and Marcus, 2000).

오류의 종류에 따라 정제의 방법도 달라지겠으나 어떤 오류든 수작업에 의한 오류의 탐지와 수정은 상당량의 인력과 시간과 비용을 필요로 한다(Maletic and Marcus, 2000). 분류학적인 오류에 대한 정제작업으로 수작업으로 표본을 동정하는 것은 매우 많은 시간과 노동력을 필요로 하는 작업이며 그 자체로서 오류의 가능성을 내포하고 있으나 현재로서는 표본의 동정은 100% 전문가 또는 준전문가의 수작업에 의지할 수 밖에 없는 상황이다(Maletic and Marcus, 2000). 자료정제를 자동화로 해결해 보고자 하는 노력들이 따랐으나 자동화된 정제방법에도 문제나 한계가 있으므로 다양한 정제 방법에 따른 장점과 문제점을 파악하고 어디에 적용하면 좋을지에 대한 연구가 있었다(Maletic and Marcus, 2000). 종자식물의 다양성 연구에 있어 자료 품질을 통제하고 향상시키며 자료의 전산화를 개선시킬 프로토콜에 대한 제안도 있었다(Hortal et al., 2007). 브라질 국립보전센터에서는

브라질 내 70개 표본관의 표본정보를 수집하여 표본관 공간자료의 품질을 평가하고 자료정제를 한 이후의 품질과 직접 비교해보는 시도를 하였다(Sodré et al., 2012).

분류학적인 자료의 질을 어떻게 평가할 것인가에 대한 연구도 진행되었는데 Stribling et al.(2003)은 수생대형무척추동물의 분류학적인 자료를 대상으로 분류학을 연구목적과 생산목적의 분류학적 조사의 두 분야로 나누고 분야에 따라 분류학적 동정이 어디에 중심을 두어야 하는 지와 자료의 품질에 대해 평가하고 소통하는 것의 중요성을 역설하였다(Stribling et al., 2003). 여러 연구실간 분류학적인 동정 결과를 비교하고 오동정의 원인을 파악하고 그에 따른 적절한 조치를 취했을 때 추후 동정 결과에 어떠한 영향을 미치는지에 대한 연구도 있었다(Stribling et al., 2008)

GBIF는 일차생물종정보의 정제에 대한 지속적인 노력을 해왔고 그 결과로 Chapman(2005)이 데이터 정제의 원칙과 방법에 대한 연구에서 자료정제의 정의와 방법론, 오류의 정의와 종류, 오류의 점검 방법 및 지리참조연산(geo-referencing)에 대해 자세히 설명하고 있다. 자료품질의 원칙에 대한 연구를 통해 박물관이나 표본관에서 디지털정보를 제공할 때 자료품질의 어떤 부분이 확인되고 관리되어야 하는지도 정리하였다(Chapman, 2005).

분류학적 정확성과 정밀성은 세 가지 유형을 가진다. 첫 번째는 정확하고 정밀한 경우이며, 두 번째는 정밀하나 정확하지 않은 경우로 검색키나 형태적 유사성에 의해 착각하여 분류하거나 명명적인 문제이고

세 번째는 가장 피하고 싶은 경우로 정확하지도 정밀하지도 않은 경우이며 이런 유형은 역량이나 운영 자체에 문제가 있음을 보여준다고 할 수 있다(Stribling et al., 2003)

GIS프로그램의 발달 덕분에 누구나 분포도를 그리는 것처럼 좌표 정보를 시각화 하거나 공간정보를 분석하는 것이 가능해졌다(Sodré et al., 2012). 그러나 공간적 오류에 대한 인식이 없거나 불확실한 데이터를 사용하면 이는 오류가 있는 결과를 낳게 되며(Chapman, 1999), 잘못된 정보를 제공하고 올바르지 못한 환경적인 결정을 내리게 한다던가 비용을 증가시킬 수 있다(Chapman, 2005). 공간자료를 효과적으로 사용하기 위한 가장 대표적인 방법은 지리참조연산으로 이는 공간 정보를 점검하고 표본 기록에 좌표를 임의로 부여하는 작업으로 표본을 보다 가치 있게 해 준다(Murphey et al., 2004). 지리참조연산을 돕기 위한 많은 온라인이나 독립형 자동화 프로그램들이 개발되고 사용되고 있는데 BioGeoMancer, GeoLoc-CRIA, GEOLocate 등이 대표적인 프로그램들이다(Chapman, 2005). 한편, Murphey et al.(2004)는 연구를 통해 자동화된 지리참조연산 방법보다는 컴퓨터의 도움을 받은 수작업에 의한 지리참조연산이 보다 정확하고 정밀하며 자동화 프로그램의 사전작업 시간을 고려한다면 시간이 꼭 더 걸리는 것은 아니라고 하였다.

자료의 오류나 정제에 대해서는 국내에서는 논문의 형태 보다는 과제로서 연구가 진행된 바 있는데 2006년도에 한국과학기술정보연구원에서 Chapman(2005)의 자료정제의 원칙과 방법을 번역하였고

2014년도에 국립수목원 산림생물표본관에서는 식물표본 자료 약 35만 점에 대한 자료를 개괄적으로 살펴보고 채집계획 수립하고자 하였다(Kim, 2014). 이 보고서에 따르면 식물표본 자료에서 채집년도의 부재가 8%, 좌표정보의 부재가 27%, 고도정보의 부재는 30%로 나타났다. 또한 채집지의 경우 강원, 전남, 경기 지역에 채집이 집중되는 편향을 보였다. 2015년도에는 산림생물표본관 소장 수목표본 10만 점(중점 재동정 분류군 11과)을 대상으로 한 재동정 사업도 진행되었는데(Chung et al, 2015) 참나무과의 49.5%, 물푸레나무과 43.3%, 노박덩굴과 25.7%, 자작나무과 14.9% 정도의 분류학적 오류를 찾아내고 주석을 달았다.

제 3 장 재료 및 방법

1. 연구 대상

본 연구의 대상은 산림청 국립수목원 산림생물표본관(Korea National Arboretum, KH)과 환경부 국립생물자원관(National Institute of Biological Resources, KB)에 소장된 식물표본과 표본정보(DB)이다.

먼저 국립수목원 물푸레나무속(*Fraxinus*) 1,158점과 수수꽃다리속(*Syringa*) 704점을 대상으로 데이터정제와 관련된 오류의 현황을 Chapman(2005)의 데이터정제의 원칙과 방법론에 따라 명명법적, 분류학적, 공간적 오류의 차원 및 관리의 관점에서 살펴보았다.

다음으로 분류학적인 오류, 즉 오동정에 대한 점검과 분석을 위해 총 17,517점의 표본을 식별, 분석하였다. 조사는 1차와 2차로 나누어 진행하였고 1차 조사에서는 물푸레나무과(Oleaceae) 전체를 확인하였다(산림생물표본관 3,194점, 국립생물자원관 4,147점). 물푸레나무과는 기존에 분포지 및 분류군의 특성에 대해 연구가 되어 있으며 분포에 있어 전국, 백두대간, 남부지역, 한정된 지역에 분포하는 분류군을 고르게 가지고 있고 1속 1종, 1속 2종, 1속 3종, 1속 4종 이상의 분류군도 고르게 가지고 있다. 또한 전체 표본의 수가 7,300점 정도로 다른 주요 과들에 비교했을 때 전수조사하기에 적절하다고 판단하였다. 2차 조사 대상은 오동정의 경향성 분석을 위해 분류군 내의 유사 종의 숫자, 분포지역 등 분석 기준에 따라 분류군을 추가하였으며 전수조사를 기본으로 하

였다(10개과, 산림생물표본관 6,038점, 국립생물자원관 4,138점). 자료의 관리와 분석을 위해 국립수목원과 국립생물자원관에서 엑셀 형식으로 제공받은 일차종발생정보(표본정보)를 DB형식으로 전환하고 브람스의 메인DB로 저장하는 방식으로 별도의 DB를 구축하였다.

2. 연구 방법

DB는 영국 옥스퍼드 대학에서 개발하여 사용중인 BRAHMS (Botanical Research And Herbarium Management System) ver. 7.9를 사용하였다. BRAHMS의 메인DB는 자료 분석의 기반임과 동시에 자료정제(data cleaning)에 이용하였다.

제공받은 자료로 구축한 DB는 표본관을 방문하여 표본을 하나씩 대조하면서 DB와 실제 표본간의 차이를 점검하였으며 재동정 작업은 대한식물도감(1982)과 한반도의 수목 필드가이드(2012), 한국의 나무(2011)와 한국의 수목 사이트(http://florakorea.myspecies.info/en)를 참고하였다. 모든 표본을 직접 동정하고 오동정된 표본은 주석을 달았으며 엑셀과 브람스를 이용하여 정보를 수정하고 정리하였다.

분류학적 정밀성(precision)의 분석은 Stribling et al(2003, 2008)의 분석방법을 따랐는데 정밀성은 같은 대상에 대해 다른 방법으로 측정하였을 때의 가까운 정도, 즉 반복성에 대한 측정을 말한다. 정밀성을 측정하는 방법은 크게 두 가지로 1) 2가지 이상의 방법으로 같은 대상에 대해 측정 하는 것과 2) 하나의 방법으로 반복적으로 측정하는 것이다. 분류학적으로 대입시키면 1) 2곳 이상의 분류학자(혹은 분류학 연구실)에서 같은 표본에 대해 동정하는 것과 2) 동일한 분류학자(혹은 분류학 연구실)가 반복적으로 같은 표본에 대해 동정하는 것으로 정리해 볼 수 있다. Chapman(2005)는 동정에 있어

정밀성을 표본을 어떤 수준으로 동정하였는가로 종(species) 차원에서의 동정은 과(family) 차원에서의 동정보다 정밀한 것으로 보기도 하였는데 이는 Stribling et al(2003, 2008)의 기준에 따르면 정확성으로 볼 수 있다.

본 논문에서는 Stribling et al.(2003, 2008)의 정의에 따라 2015년도에 국립수목원에서 실시한 재동정 사업의 내용과 본 연구를 위해 2017년도에 재동정 한 결과를 통해 정밀성을 측정하였다.

숫자의 합의 차를 나타내는 PDE(percentage difference in enumeration)은 다음과 같은 식에 따라 계산하였다. 여기서 $n1$은 2015년도에 재동정한 표본 수, $n2$는 2017년도에 재동정한 표본 수이다.

$$PDE = \frac{|n1 - n2|}{n1 + n2} * 100$$

두 번의 재동정의 분류적인 불일치 정도(PTD, % taxonomic disagreement)는 다음의 식에 따라 계산하였다. $Comp_{pos}$는 동정 결과에서 서로 일치한 수이다. N은 2015년도와 2017년도 표본수의 평균으로 하였다.

$$PTD = \left[1 - \left(\frac{comp_{pos}}{N}\right)\right] * 100$$

분포도를 통한 오류의 탐지는 동정한 자료를 근거로 DB상의 표본의 좌표와 지명 정보를 이용하여 실시하였다. 1945년 이전 채집자료와 한국의 수목 사이트(http://florakorea.myspecies.info/en)를 기준했을 때의 분류군의 분포와 표본 DB의 분포를 좌표와 행정구역으로 비교하여 분포지가 아닌 곳에서 채집한 것으로 생각되는 자료를 골라내고 분포도

작업을 통해 그 중 실제 오동정이 얼마나 되었는지를 확인하였다. 분포도 및 분포정보를 위한 GIS(geographic information system)는 DIVA-GIS ver.7.5를 활용하여 BRAHMS DB를 엑셀 형식으로 추출한 후 DIVA-GIS의 shp파일로 변환 후 사용하였다.

제 4 장 결 과

1. 국립수목원(KH) 표본 DB 현황 및 품질

1.1. 표본 DB 현황

표본자료의 정리와 확인에 가장 기본이 되는 중요한 정보는 표본고유번호(accession번호)와 바코드번호인데 이 체계에 일관성이 부족했다. KHB로 시작되는 원래의 표본고유번호(관리번호)가 제대로 입력되어 있는 표본은 물푸레나무속은 73.9%, 수수꽃다리속은 79.1%였다. KHB가 아닌 KNKA로 시작되는 표본고유번호가 물푸레나무속은 26.1%, 수수꽃다리속에서는 20.9%를 차지하였다. 이 중에서도 KHB로 시작되는 번호가 표본고유번호 입력란이 아닌 바코드 입력란에 입력되어 있는 자료들도 있었다. 정리해보면 표본고유번호에 있어 1) KHB형식이면서 바코드란에 동일한 KHB형식의 번호가 입력되어 있는 자료, 2) KNKA형식이면서 바코드란에 KHB형식의 번호가 입력된 자료, 3) KNKA형식이면서 바코드란은 비워진 자료 등 3가지 유형으로 구분할 수 있었다(Table 1). 한편, KNKA 형식의 표본고유번호의 경우 동일한 번호가 2-3개씩, 많게는 10개씩 반복되는 경향이 있어 DB의 참조무결성에 문제가 있었다.

Table 1. Three types of accession number and barcode number.

Accession No.	KHB type No.	KNKA type No.	
Barcode No.	Same as accession No.	KHB type No.	None
Fraxinus	881 (73.9%)	43(3.6%)	268(22.5%)
Syringa	508 (79.1%)	23(3.6%)	111(17.3%)

표본 라벨 기본정보의 부재 현황은 다음과 같았다. 물푸레나무속의 경우 DB 1,192개 중 채집자 정보가 없는 표본은 256점(20.0%)이었고 날짜에 있어 채집년도가 없는 표본은 250점(19.5%), 채집월이 없는 표본은 252점(19.7%), 채집일이 없는 표본은 263점(20.5%)이었다. 수수꽃다리속의 경우 채집자 정보가 없는 표본은 134점(19.4%)이었고 날짜에 있어 채집 년도가 없는 표본은 122점(17.3%), 채집 월이 없는 표본은 123점(17.5%), 채집일이 없는 표본은 125점(17.8%)이었다.

DB와 표본을 일대일로 대조해본 결과 물푸레나무속 DB 1,192개 중 외국 표본 34개를 제외한 1,158개를 대상으로 하였을 때, 452개의 자료(37.9%)에 대한 표본을 찾을 수 없었다. 이 중에는 표본고유번호도, 바코드번호도 없었던 268점(22.0%)도 포함되어 있기 때문에 DB상에 있으나 확인할 수 없는 표본이, 표본은 있으나 DB에는 없는 표본과 일부 일치할 수도 있다. 그러나 대부분이 채집일(87.1%)이나 채집자(87.8%)에 대한 정보도 없었기 때문에 어느 표본인지 확인할 수 없어 날짜나 장소로 확인이 가능한 것만 확인하여 표본이 있는 것으로 처리하였다. 수수꽃다리속의 경우 DB 642개 중 185개의 자료(28.8%)에 대한 표본을 찾을 수 없었다.

DB가 누락된 표본 수(DB는 없고 표본만 있는 경우)는 물푸레나무[*Fraxinus chinensis* Roxb. var. *rhynchophylla* (Hance) Hemsl.] 131점, 쇠물푸레나무(*Fraxinus sieboldiana* Blume) 157점, 들메나무(*Fraxinus mandshurica* Rupr.) 37점, 물들메나무(*Fraxinus chiisanensis* Nakai) 37점, 구주물푸레나무(*Fraxinus excelsior* L.) 16점으로 총 378점의 표본에 대한 DB를 확인할 수 없었다. 기타 물푸레나무속 종들의 현황을 정리해보자면, 좀쇠물푸레(*Fraxinus sieboldiana* var. *angusta* Blume)의 경우 DB에 62개의 자료가 있고 이 중 22점의 표본은 2015년도에 쇠물푸레로 재동정 되었으며 4점은 그대로 좀쇠물푸레 폴더에 남아 있었고 나머지는 찾을 수 없었다. 광릉물푸레[*Fraxinus rhynchophylla* var. *densata* (Nakai) Y.N.Lee]의 경우는 DB상에 10개의 자료가 확인되었으며 학명은 *Fraxinus chinensis* var. *rhynchophylla*로, 국명은 광릉물푸레로 되어 있었다. 표본은 4점만 확인할 수 있었고 표본의 라벨에는 학명이 제대로 되어 있었다. 긴잎쇠물푸레는 DB상 학명이 *Fraxinus sieboldiana*로 되어 있고 10개 중 2점의 표본을 확인할 수 있었다. 구주물푸레나무는 DB 6개 중 확인된 표본은 1점이었고 DB에 없으나 16점의 표본을 추가로 확인할 수 있었다. 수수꽃다리속의 경우 털개회나무[*Syringa pubescens* subsp. *patula* (Palibin) M.C. Chang & X.L. Chen]는 241점, 개회나무[*Syringa reticulate* (Blume) H. Hara] 60점, 꽃개회나무(*Syringa wolfii* C.K. Schneid) 34점 등 총 335 점의 표본에 대한 DB를 확인할 수 없었다.

1.2. 표본 DB 품질

1.2.1. 명명법적 문제

이명인 좀쇠물푸레(*Fraxinus sieboldiana* var. *angusta* Blume)와 학명 확인이 안 되는 긴잎쇠물푸레나무(DB상은 *Fraxinus sieboldiana*)가 분류군으로 관리되고 있었다. 좀쇠물푸레나무는 2015년 재동정 당시 쇠물푸레나무로 합쳐지면서 이명처리 되었다고 하나 일부 표본은 아직 폴더에 그대로 남아 있었다.

물들메나무를 칭하던 긴잎물푸레나무는 DB상 학명은 쇠물푸레나무(*Fraxinus sieboldiana*)로 되어 있는데 긴잎쇠물푸레나무의 오타이거나 작업자의 실수로 추측된다.

DB로 확인하였을 때 물푸레나무속의 경우 정명이 61.1%, 이명이 37.6%, 비합법명 등의 이유로 처리되지 않은 분류군이 1.2%, 불확실한 분류군이 0.1%였으며 수수꽃다리속의 경우는 정명이 27.1%, 이명이 70.3%, 불확실한 분류군은 2.7%였다. 수수꽃다리속의 경우 이명 처리되었으나 DB에 반영되지 않은 분류군이 많았다.

1.2.2. 분류학적 정밀성

본 논문에서는 Stribling et al.(2003, 2008)에 따라 2015년도에 국립수목원에서 실시한 재동정 사업의 내용과 본 연구를 위해 2017년도에 재동정 한 결과를 통해 정밀성을 측정하였다.

먼저 동정한 표본의 수를 비교함으로써 자료의 질을 살펴보았다. 숫자의 합의 차를 나타내는 PDE(percentage difference in enumeration)는 여러 분류군의 식물을 주고 분류학자(연구실) 간 어느 종을 어느 분류군으로 얼만큼씩 분류하는지에 대한 정확성을 보기 위한 것이나 본 논문에서는 표본의 관리 상황을 점검하는 차원으로 2015년도에 재동정한 분류군별 숫자와 2017년도의 재동정 숫자와 결과를 비교함으로써 그 차이를 보고자 하였다(두 재동정 작업은 동일한 시기의 표본에 대한 전수 조사였으므로 원칙적으로는 재동정한 표본의 숫자가 같아야 한다.). 그 결과, 꽃개회나무의 표본 수 간의 차이가 12.9%로 가장 큰 것으로 나타났다(Figure 1).

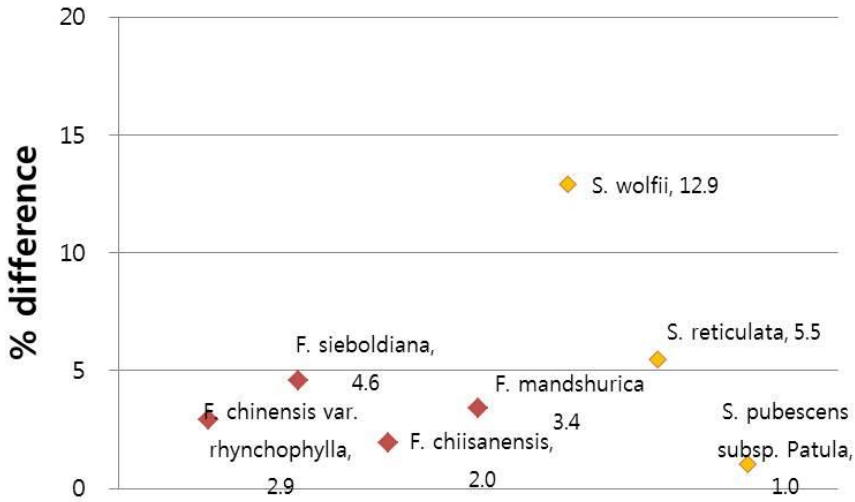


Figure 1. Comparison of percentage difference in enumeration(PDE) for specimens re-identified by two laboratories. PDE shows the difference between year 2015 and 2017. In case of *Syringa wolffi*, 118 specimens were re-identified in 2015 while 153 specimens in 2017 so that the difference was 12.9%.

재동정의 분류학적 불일치 정도(PTD, % taxonomic disagreement)는 들메나무와 물들메나무에서 각각 48.3%와 39.2%로 높게 나타났다(Table 2-3, Figure 2).

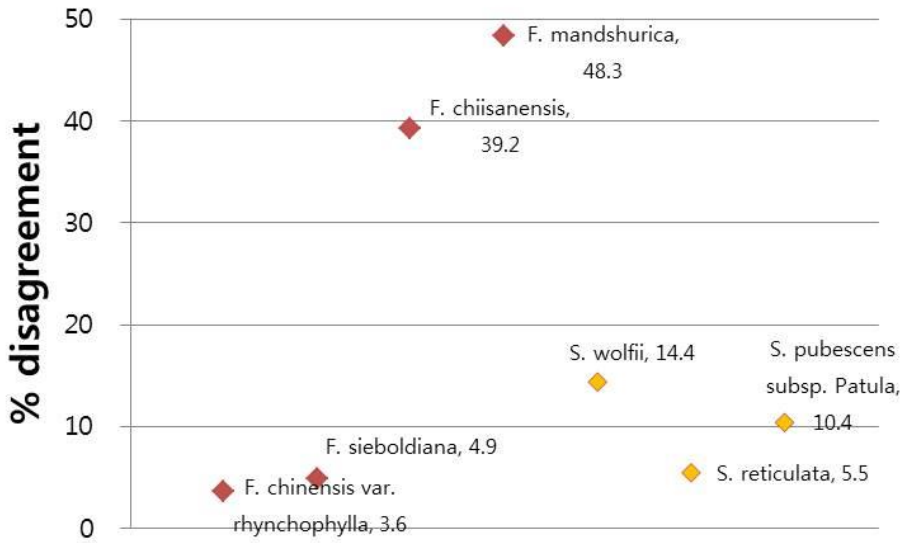


Figure 2. Comparison of percentage taxonomic disagreement (PTD) for specimens re-identified by two laboratories. PTD shows the disagreement rates between two re-identification results. The results of *Fraxinus mandshurica* and *Fraxinus chiisanensis* were considerably higher than other taxa.

Table 2. Comparisons of re-identification results of *Fraxinus* showing counts of agreements. Target taxonomic level is species. In case of *Fraxinus mandshurica*, less than half of the specimens(38 specimens) were agreed.

Identification	2015	2017	No. agreements
Fraxinus chinenesis* var. *rhynchopylla	447	471	444
F.sieboldiana	-	1	-
F.chiisanensis	-	1	-
Other taxon	-	1	-
F. sieboldiana	533	564	531
F. chinensis var. *rhynchopylla*	-	14	-
Other taxa	-	6	-
F. chiisanensis	39	28	28
F. chinensis var. *rhynchopylla*	3	24	3
F. mandshurica	8	-	-
F. mandshurica	63	29	29
Fraxinus chinenesis var. *rhynchopylla*	7	30	7
F.sieboldiana	2	2	2
F.chiisanensis	-	14	-
Other taxon	1	1	-

Table 3. Comparisons of re-identification results of *Syringa* showing counts of agreements. Target taxonomic level is species.

Identification	2015	2017	No. agreements
Syringa wolfii	98	134	98
S. reticulata	3	3	3
S. pubescens subsp. *patula*	17	15	15
Other taxon	–	1	–
S. reticulata	181	189	181
Syringa wolfii	–	5	–
S. pubescens subsp. *patula*	–	7	–
Other taxa	–	1	–
S. pubescens* subsp. *patula	96	86	86
Syringa wolfii	–	3	–
S. reticulata	1	5	–
Other taxa	–	1	–

1.2.3. 공간 정보의 충실도

Wieczorek et al(2004)은 라벨상의 공간정보의 유형(geocode type)이 9개로 구분되고 결정된다고 하였다. 이에 따라 물푸레나무속과 수수꽃다리속의 공간정보 현황을 정리해 보면, 지명만 표시된 것이 49.1%, 좌표정보까지 있는 자료는 48.3%였으며 나머지는 모호하거나(0.9%) 장소를 추정하기 어려운 넓은 범위로만 써있는 경우(0.1%)였다. 공간 정보가 아예 없는 경우(1.6%)도 있어 이를 유형으로 새롭게 추가하였다. 수수꽃다리속의 경우도 결과가 비슷하였다(Table 4).

Table 4. Ten categories of geocode type. Type of 'No data' is added to Wieczorek's nine categories because there are labels with no spatial data.

Type	Example	No.(%) of specimen	
		Fraxinus	*Syringa*
Dubious	Myogol(묘골)	11(0.9%)	3(0.5%)
Cannot be located	Seoul	1 (0.1%)	5(0.8%)
Demonstrably inaccurate		−	−
Coordinates	Latitude, Longitude	559(48.3%)	366(57%)
Named place	Mt.Jiri	569(49.1%)	261(40.7%)
Offset		−	−
Offset along a path		−	−
Offset in orthogonal directions		−	−
Offset at a heading		−	−
No data *		18(1.6%)	7(1.1%)
	TOTAL	1,158	649

* This type is added.

2. 국립수목원(KH)과 국립생물자원관(KB) 표본 DB 의 공간적 및 분류학적 오류

공간적 오류와 분류학적 오류의 현황에서는 주로 분류학적인 오류에 초점을 맞추었다. 분류학적인 오류, 즉 오동정은 표본을 열람한 국내 연구자들이 가장 많이 접하는 오류이며 오류의 정도에 대해서도 다양한 의견을 가지고 추측하고 있으나 그간 확인되고 알려진 바가 없다.

2.1. 공간적 오류

물푸레나무과을 대상으로 국립수목원(KH)과 국립생물자원관(KB)의 공간적인 오류를 분석한 결과, 국립수목원의 경우 채집 년도가 없는 자료가 17%, 좌표정보가 없는 자료는 47%, 일부 정보만 있는 등 불명확한 경우가 2%였으며 고도정보는 59%가 없었다. 국립생물자원관의 경우는 채집 년도는 거의 100% 기록되어 있었고, 좌표정보의 경우 4%, 고도정보는 39%가 누락되어 있었다.

두 기관의 결과에 차이가 있는 것은 과거 표본일수록 라벨에 공간정보가 누락되는 경향이 영향을 주었다고 판단된다. 국립수목원 산림생물표본관의 경우는 GPS 시스템이 본격적으로 도입되었던 2000년도 이전의 데이터가 많고 이들 중 다수가 입력이 미비한 자료들이다(물푸레나무과의 경우 약 26.0%). 국립생물자원관 수장고(표본관)의 경우는

2000년 이전 자료가 10.8%정도로 대부분 2000년 이후 표본이었다. 고도정보의 경우에는 두 기관 모두 높은 누락률을 보였다.

2.2. 분류학적 오류 (오동정)

2.2.1. 분류학적 오류의 현황과 경향성

분류학적 오류라 할 수 있는 오동정의 정도에 있어서는 분류군 별로 차이가 컸다. 우선 물푸레나무과를 살펴보면 물푸레나무과 재동정 결과 들메나무의 오동정률이 66.1%로 가장 높았고 다음으로 왕쥐똥나무(*Ligustrum ovalifolium* Hassk.)가 52.2%, 상동잎쥐똥나무(*Ligustrum quihoui* Carrière)와 제주광나무(*Ligustrum lucidum* W.T. Aiton)가 각 48.2%와 44.8%, 산동쥐똥나무[*Ligustrum leucanthum* (S.Moore) P. S. Green] 33.3% 순으로 높았는데 대부분의 쥐똥나무속(*Ligustrum*)이 높은 오동정률을 보였다. 이팝나무(*Chionanthus retusus* Lindl. & Paxton)의 오동정률은 2.0%였고, 박달목서(*Osamanthus insularis* Koidz.)는 2.5%, 섬쥐똥나무(*Ligustrum foliosum* Nakai)는 2.8%였고, 미선나무(*Abeliophyllum distichum* Nakai)는 오동정이 없었다(Figure 3).

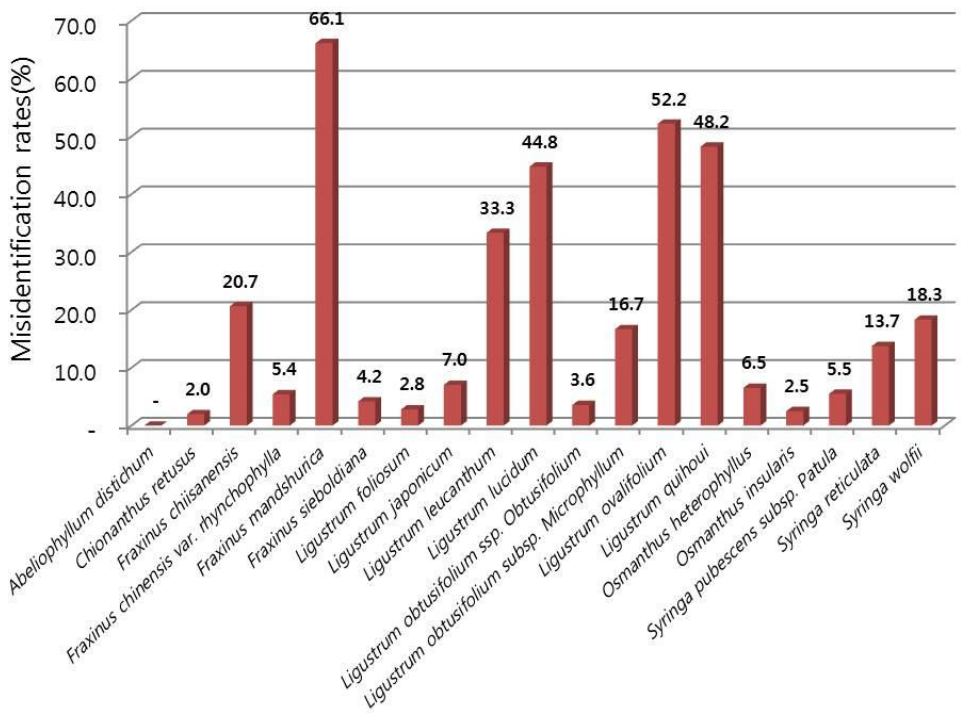


Figure 3. Misidentification rates of family Oleaceae. The rates were substantially different among species.

기타 분류군을 살펴보면 참개암나무(*Corylus siboldiana* Blume)의 오동정률이 67.07%로 가장 높았고 말채나무(*Cornus walteri* Wangerin), 새우나무(*Ostrya japonica* Sarg.) 등도 각각 34.7%, 27.1%로 오동정이 많았다. 1속 1종의 경우 오동정이 없거나 오동정률이 매우 낮았는데 고추나무(*Staphylea bumalda* DC.)는 0.8%, 푸조나무[*Aphananthe aspera* (Thunb.) Planch.]는 0.6%였다. 새우나무의 경우만 이례적으로 27.1%를 보였는데 이는 국립수목원 표본 47점 중 산림과학원으로부터 기증받은 표본 16점이 모두 개서어나무[*Carpinus*

tschonoskii (Siebold & Zucc.) Maxim]로 오동정 되었기 때문으로 편향(bias)이 있다고 보았다(Figure 4). 등칡(*Aristolochia manshurien-sis* Kom), 조도만두나무(*Glochidion chodoense* J.S. Lee & H.T. Im), 후박나무(*Machilus thunbergii* Siebold & Zucc. ex Meisn.), 상산(*Orixa japonica* Thunb.), 다정큼나무[*Rhaphiolepis indica* var. *umbellate* (Thunb.)H. Ohashi]는 오동정이 없었다.

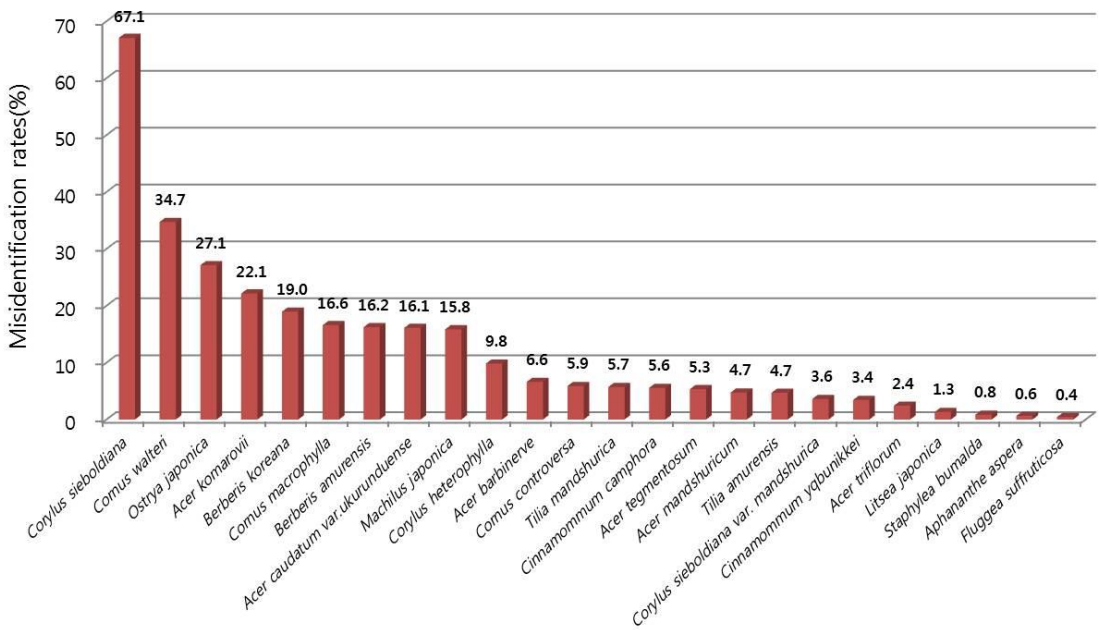


Figure 4. Misidentification rates of other taxa. The rates were substantially different among species.

조사대상 전체(49종) 17,517점의 표본 중 1,818점이 오동정으로 오동정률 평균은 10.4%였다. 오동정률은 주로 5%에서 25% 정도였으

나 오동정이 없는 것부터 67.1%까지 분류군 간의 편차가 크게 나타났다(Figure 5).

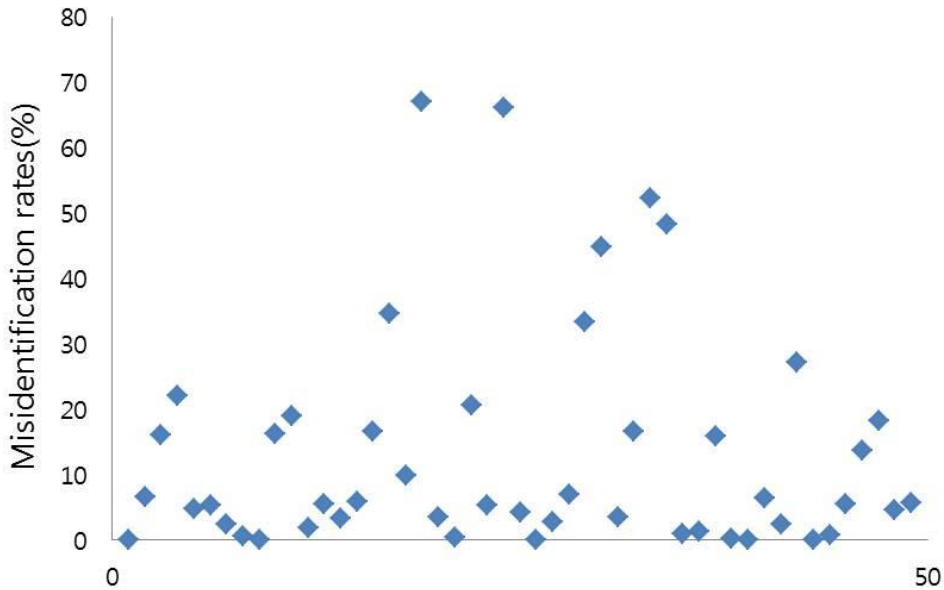


Figure 5. Misidentification rates for all re-identified species(49 species). The misidentification rates were mostly between 5 and 25%.

속 별로 살펴보면 새우나무속(*Ostrya*)의 오동정률이 27.1%로 가장 높으나 위에서 언급한대로 표본수가 적고 오동정 표본이 모두 특정 채집기관에서 기증된 표본이어서 분류군의 오동정률을 보여준다고 하기는 어려웠다. 개암나무속(*Corylus*), 매자나무속(*Berberis*), 층층나무속(*Cornus*)이 15% 이상의 높은 오동정률을 보였으며 단풍나무속(*Acer*), 물푸레나무속, 쥐똥나무속(*Ligustrum*), 수수꽃다리속도 10% 전후의 오동정률을 보였다. 그러나 속 별로 살펴보았을 때 종 간 차이가 물푸레나무속의 경우 4종 간 3.9%부터 64%까지, 쥐똥나무속은 8종 간 2.8%부

터 52.2%까지로 차이가 워낙 크기 때문에 분류군과 오동정률 예측에 있어 상관관계는 없는 것으로 보인다.

속 내 종 수에 따른 오동정률은 경향성이 있었는데 1속 내 1종만 있는 그룹의 오동정률은 1.3%, 1속 2종은 6.5%였고, 1속 3종은 18.3%로 가장 높았다. 1속 4종 이상의 경우 9.7%로 오동정률이 속 내 종수에 따라 계속 증가하는 것은 아니고 3종 정도의 유사한 종이 있을 때 가장 높다는 것을 알 수 있었다. 상자그림으로 살펴보면 1속 내 4종 이상이 있는 그룹의 경우 종 간 편차가 보다 컸다(Figure 6).

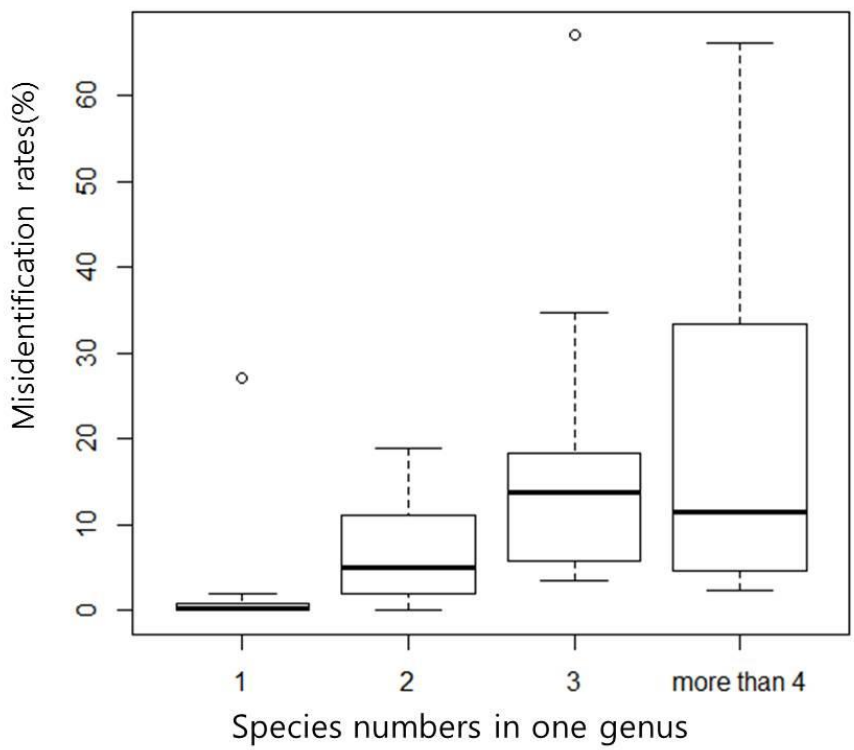


Figure 6. Misidentification rates by the number of species in one genus. The rates became higher with increasing species numbers in one genus.

분포지별 오류의 현황에서는 한정된 지역, 혹은 좁은 지역에 분포하는 종의 오동정률이 40.2%로 가장 높았고, 백두대간 분포가 15.3%, 남부 서해안 분포 6.4%, 전국분포 6.3%를 보여 분포지별 경향성을 볼 수 있었는데 한정된 지역에 분포하는 그룹의 경우 종간 편차가 큰 것을 볼 수 있었다(Figure 7).

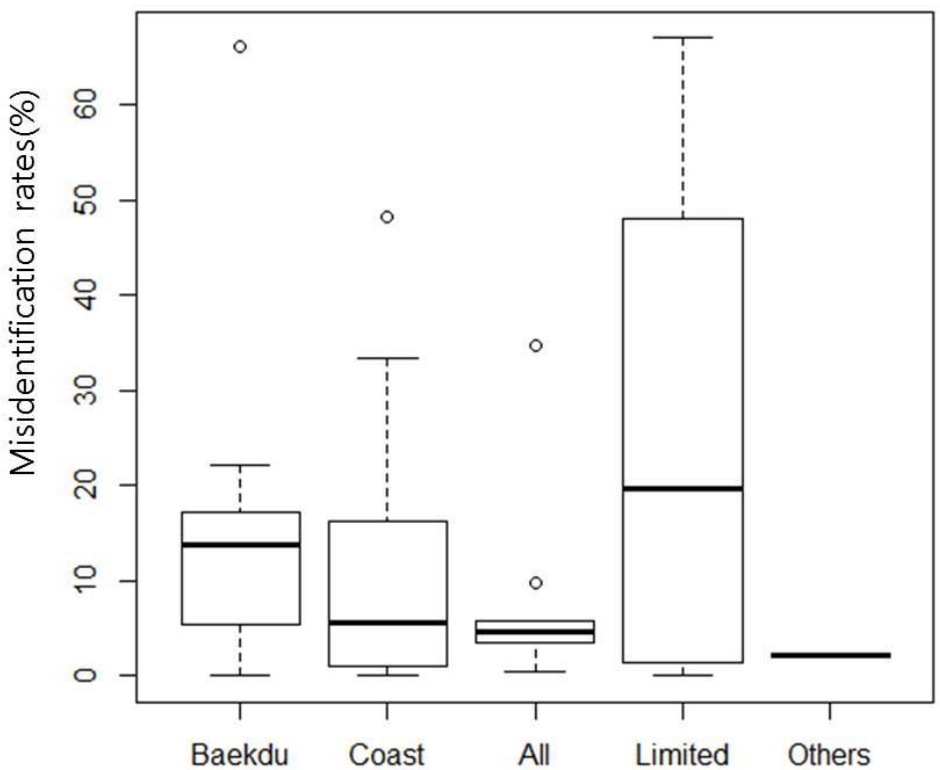


Figure 7. Misidentification rates by natural distribution area of species. The misidentification rates were higher for the species distributed in Baekdudaegan and limited area.

2.2.2. 분류학적 오류의 유형

오동정의 유형은 정확성(accuracy)와 관련된 것으로 기준표본이나 기재문 등 동정의 기준으로 삼는 것에 얼마나 근접하여 동정을 하였는지를 살펴볼 수 있다(Stribling et al., 2003; 2008). 본 연구에서는 오동정이 동일한 분류군 내에서 발생하였다면 정확성이 높은 것으로 판단하였는데 전체 오동정의 91.5%가 같은 속 내의 다른 종으로 동정된 것이었으며(Figure 8) 1.3%는 같은 과 내의 오동정으로(Figure 9) 잎과 꽃, 열매 등 전체적인 형태의 유사성에 의한 오동정이었다. 7.2%는 전혀 다른 분류군으로 동정된 경우로 주로 잎이나 꽃, 열매의 일부 형태가 유사한 경우였으나(Figure 10) 국명의 유사성으로 인해 라벨이 잘못 제작되는 등의 단순한 착오로 인한 오류도 있었다(Figure 11).

Table 5. Type and percentage of misidentified specimens.

Within same genus	Within same Family	Between other taxon	Total
1,674 (91.5%)	24 (1.3%)	131 (7.2%)	1,829

Figure 8. *Corylus sieboldiana* Blume var. *mandshurica* (Maxim.) C.K. Schneid, misidentified with *Corylus sieboldiana* Blume.

Figure 9. *Alnus incana* (L.) Medik. subsp. *hirsuta*, misidentified with *Corylus heterophylla* Fisch. ex Trautv.

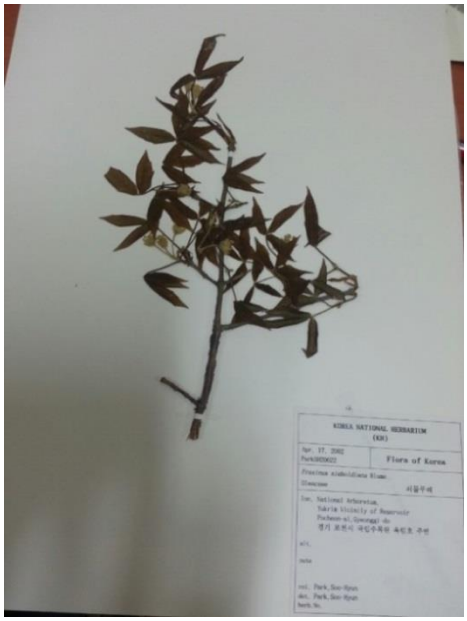


Figure 10. *Fraxinus sieboldiana* Blume misidentified with *Acer triflorum* Kom.

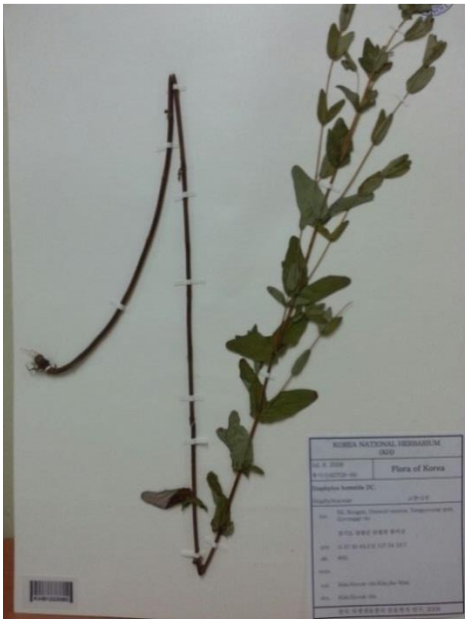


Figure 11. *Hypericum erectum* Thunb., misidentified with *Staphylea bumalda* DC. Korean name(local name) of these species are very similar("고추나무" and "고추나물").

또한 같은 속에 속하면서 형태적으로 유사한데 한 종이 더 흔하고 잘 알려져 있고 다른 종이 상대적으로 덜 흔하고 덜 알려져 있을 때, 더 흔하고 더 알려진 종을 덜 흔하고 덜 알려진 종으로 동정하는 비율이 높았다. 예로 들면, 물푸레나무과에서 들메나무 중 41.5%가 물푸레나무인데 들메나무로 동정한 것이었으나 들메나무를 물푸레나무로 동정한 비율은 물푸레나무의 0.4%였다(Figure 12). 또한 광나무를 제주광나무로 동정한 비율은 57.1%이나 제주광나무를 광나무로 동정한 것은 10.9%로 이러한 경향은 다른 분류군에서도 공통적으로 볼 수 있었다.

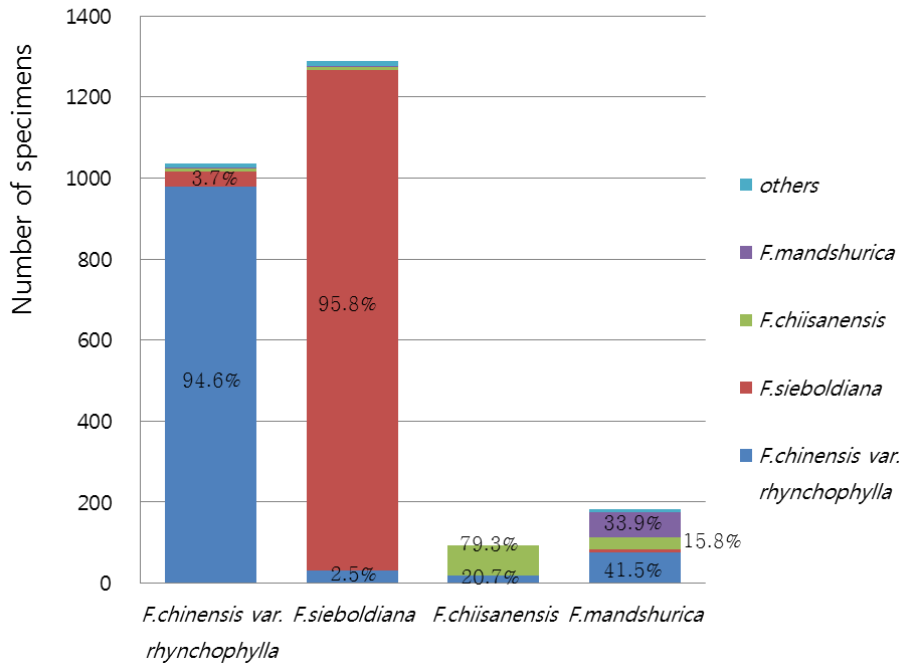


Figure 12. The difference between misidentification rates of more common and less common species. Strong tendency is shown that collectors misidentify even more when the species is common and well-known.

3. 분포도 작성을 통한 오류의 예측

3. 1. 분포도와 오류의 탐지

분포도를 작성해 보면 더욱 명확하게 오동정된 표본을 탐지할 수 있으며 좌표 입력 오류 혹은 좌표 자체가 잘못된 자료까지도 탐지할 수 있다.

부게꽃나무[*Acer caudatum* subsp. *ukurunduense* (Trautv. & C. A. Mey.) A. E. Murray], 청시닥나무, 시닥나무(*Acer komarovii* Pojark.), 복장나무(*Acer mandshuricum* Maxim.), 꽃개회나무 등은 백두대간을 중심으로 분포하는 종으로 벗어난 지역은 오동정이거나 좌표 오류임을 알 수 있다(Figure 13). 광나무(*Ligustrum japonicum* Thunb.), 푸조나무, 녹나무, 육박나무(*Cinnamomum yabunikkei* H. Ohba), 곰의말채나무(*Cornus macrophylla* Wall), 쇠물푸레나무 등은 남부 지방이나 해안을 따라 분포하고 있다(Figure 14). 개암나무, 피나무(*Tilia amurensis* Rupr.), 층층나무(*Cornus controversa* Hemsl.), 말채나무 등은 전국분포를 보여 분포를 통한 오동정의 탐지는 어려웠으나 좌표오류에 대해서는 정제가 가능했다(Figure 15). 매자나무(*Berberis koreana* Palib.), 미선나무, 물들메나무, 조도만두나무, 섬쥐똥나무 등은 한정된 지역 혹은 좁은 지역에 분포하여 분포도 작업을 통한 오동정의 탐지 및 확인이 용이하였다(Figure 16).

분포지 이외에서 표본이 확인된 종들도 있었는데 이팝나무는 경기도 및 전라도, 제주도 분포로 알려져 있었으나 충청도, 경상북도, 강원도에서도 표본이 채집되었고 복자기나무(*Acer triflorum* Kom.)는 기존에 백두대간 및 경기 북부 분포로 알려져 있었으나 경기남부와 충청남도까지 분포함을 확인하였다(Figure 17). 그러나 두 종 모두 조경수로 즐겨 쓰이는 종이므로 조경용으로 식재된 개체 영향으로 추정되는데 이런 종들은 분포를 통한 오동정의 탐지에 어려움이 있었다.

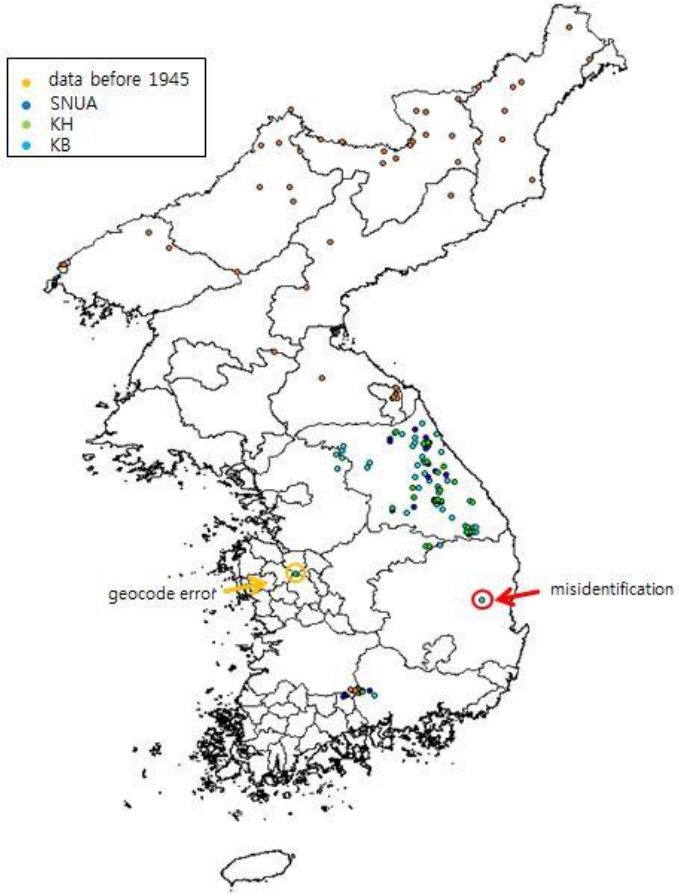


Figure 13. Distribution of *Acer caudatum* var. *ukurunduense* (Trautv.&C.A. Mey.) Rehder.

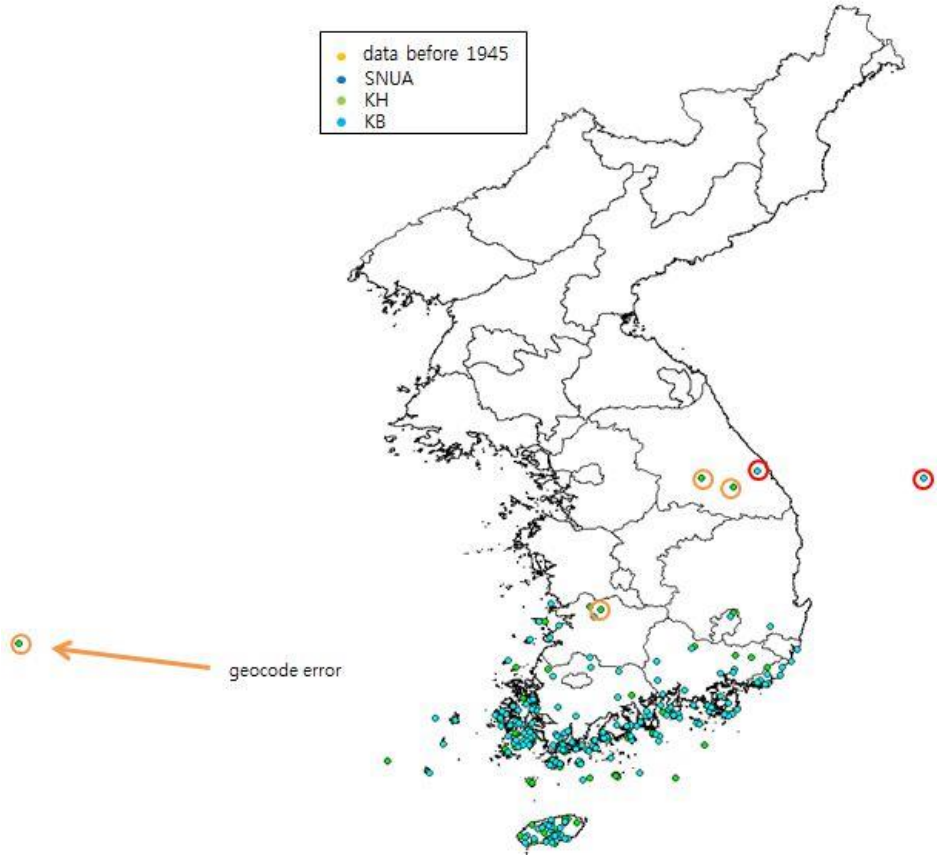


Figure 14. Distribution of *Ligustrum japonicum* Thunb.

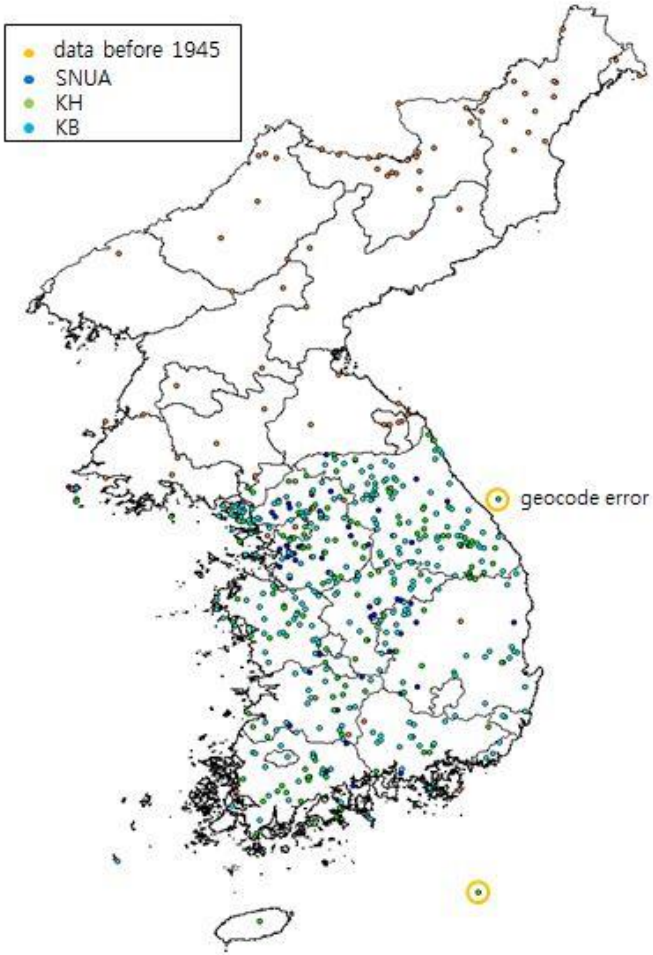


Figure 15. Distribution of *Corylus heterophylla* Fisch. Ex Trautv.

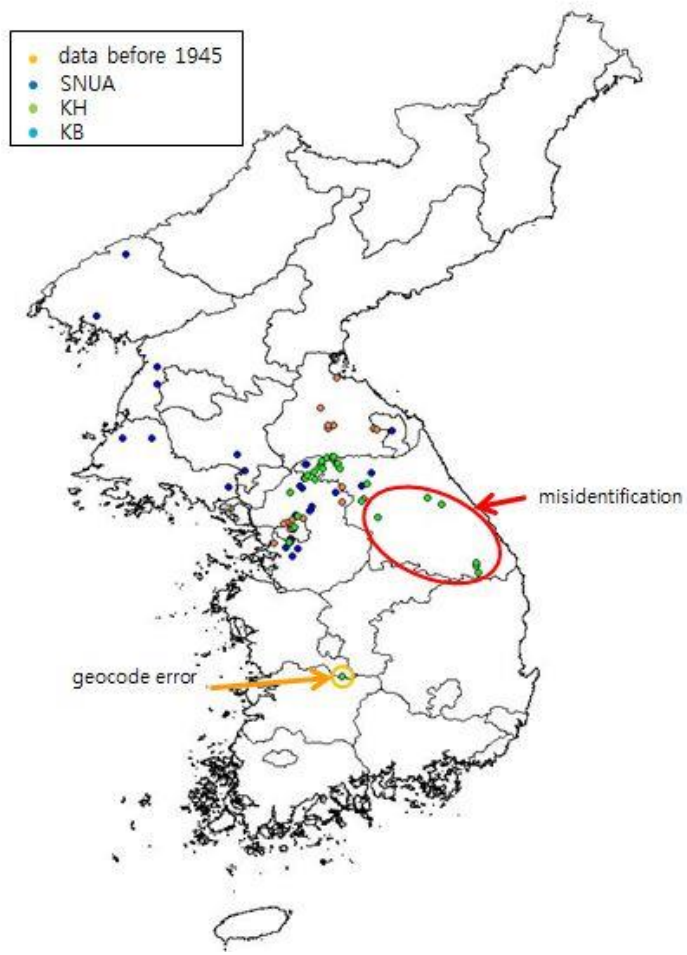


Figure 16. Distribution of *Berberis koreana* Palib.

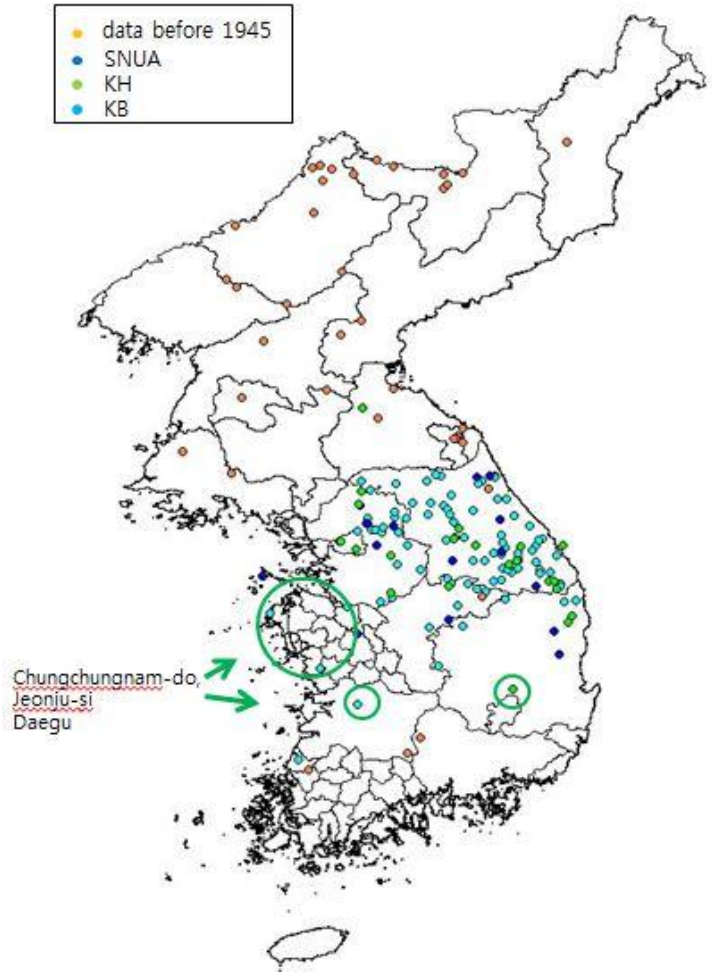


Figure 17. Distribution of *Acer triflorum* Kom.

3. 2. 종별 분포 유형과 예측의 결과

분포도를 그린 결과에 따라 분포패턴을 다섯 가지로 유형화하고 그룹별 예측률을 살펴보았다(Table 6).

조사대상 17,517점 중 오동정은 1,829점이었는데 이 중 468점의 오동정(25.6%)을 분포로 탐지할 수 있었다. 한정된 지역이나 좁은 지역에 분포하는 수종의 경우 탐지율이 높았는데 울릉도에만 분포하는 섬쥐똥나무의 경우는 100%, 참개암나무 65.0%, 물들메나무는 42.1%의 오동정을 탐지할 수 있었다. 백두대간을 중심으로 분포하는 들메나무는 64.5%, 청시닥나무(*Acer barbinerve* Maxim.) 53.3%, 꽃개회나무는 44.0%, 복장나무 40.0% 등으로 오동정을 탐지할 수 있었다.

반면, 전국 혹은 넓은 지역에 분포하는 물푸레나무, 쥐똥나무(*Ligustrum obtusifolium* Siebold & Zucc.)등의 경우는 좌표오류 이외에 오동정을 탐지하기는 어려웠다. 남부해안 지역에 분포하는 좀털쥐똥나무[*Ligustrum obtusifolium* Siebold & Zucc. subsp. *microphyllum* (Nakai) P.S. Green], 상동잎쥐똥나무(*Ligustrum quihoui* Carrière), 왕쥐똥나무, 박달목서 등의 예측률도 대부분 5% 미만으로 낮았는데 국립표본관과 같은 채집을 하는 기관들이나 학교들의 다수가 서울, 경기도권에 위치하기 때문에 남쪽에 분포하는 식물의 경우 채집을 위해 분포지를 찾아가는 경우가 많아서인 것으로 생각된다.

Table 6. Taxa were categorized into five groups on the basis of natural distribution area of the species.

Group	Group 1	Group 2	Group 3	Group 4	Group 5
Area	Baekdudaegan	Southern, western coast	All	Limited	Other
Species	*Acer barbinerve* *A.caudatum* var. *ukurunduense* *A. komarovii* *A.mandshuricum* *A.tegmentosum* *Aristolochia manshuriensis* *Berberis amurensis* *Fraxinus mandshurica* *Syringa pubescens* subsp. *patula* *S. reticulata* *S. wolfii*	*Aphananthe aspera* *Cinnamomum camphora* *C. yabunikkei* *Cornus macrophylla* *Fraxinus sieboldiana* *Ligustrum japonicum* *L.leucanthum* *L.obtusifolium* subsp. *Microphyllum* *Ligustrum quihoui* *Litsea coreana* *Litsea japonica* *Machilus japonica* *M.thunbergii* *Orixa japonica* *Osmanthus heterophyllus* *O. insularis* *Ostrya japonica* *Rhaphiolepis indica* var. *umbellata* *Tilia mandshurica*	*Cornus controversa* *C. walteri* *Corylus heterophylla* *C.sieboldiana* var. *mandshurica* *Fluggea suffruticosa* *Fraxinus chinensis* var. *rhynchophylla* *Ligustrum obtusifolium* *Staphylea bumalda* *Tilia amurensis*	*Abeliophyllum distichum* *Berberis koreana* *Corylus sieboldiana* *Fraxinus chiisanensis* *Glochidion chodoense* *Ligustrum foliosum* *L.lucidum* *L.ovalifolium*	*Acer triflorum* *Chionanthus retusus*

한편, 전국적인 분포를 보이는 종이나 분포지로 판단했을 때 의심되는 표본자료가 없었던 녹나무[*Cinnamomum campphora* (L) J. Presl], 박달목서, 센달나무(*Machilus japonica* Siebold & Zucc. ex Meisn) 등의 분류군을 제외한 24종만을 대상으로 하였을 때 예측률은 33.1%였다 (Figure 18).

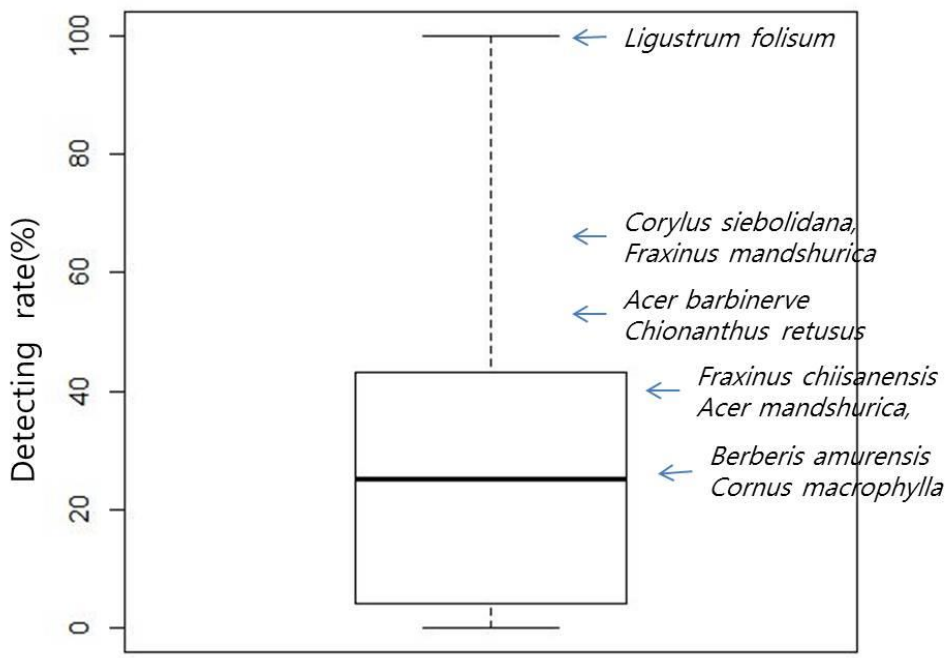


Figure 18. Detecting rates of misidentification.

제 5 장 고 찰

Redman(2001)은 문제가 있는 데이터베이스를 오염된 호수에 비유하였다. 호수에 오염된 물을 공급하는 공장은 적절하지 못한 자료를 공급하는 주체이고 오염된 물을 마시는 사람은 자료의 사용자라는 것이다. 이 비유에 따르자면 그 동안 우리나라의 상황은 호수의 크기와 유입되는 물의 양에만 신경을 썼을 뿐 어떤 품질의 물을 공급하고 마시고 있는지에 대한 관심이 부족했다.

1. 표본 자료 데이터베이스화의 의미

Haston et al.(2012)는 왕립식물원(Royal Botanic Garden)의 표본 정보의 디지털화 작업과정을 구축하면서 기존 표본작업이나 큐레이션(curation)이 반드시 구조적으로 같이 진행되어야 한다고 강조하였다. 국립수목원(KH) 물푸레나무속의 경우 DB상의 37.4%의 표본을 실제로 확인할 수 없었으며, DB에 없는 표본은 DB상의 물푸레나무속 39%에 해당하였다. 좀쇠물푸레의 경우 DB와 비교했을 때 60%이상의 표본을 찾을 수 없었다. DB상의 표본 수 1,158점과 실제 표본 수 1,146점을 숫자상으로만 비교하면 거의 오류가 없는 것으로 볼 수 있으나 실제로는 DB상의 37.4%, 실제 표본의 39.4%가 오류인 것이다. 실제로는 재동정되어 주석이 달려 있으나 DB상에는 여전히 예전 종으로 되어 있거나 주

석이 달린 표본이 어떤 것은 기존 분류군의 폴더에, 어떤 것은 재동정된 분류군의 폴더에 보관되어 있기도 했다.

기증 표본 관리의 문제점도 확인되었다. 국립수목원(KH)의 경우 산림과학원에서 기증받은 표본의 비율이 매우 높았는데 다수의 표본이 산림과학원에 식재되었던 것이며 라벨의 내용이 년도만 기재되거나 장소가 없는 표본들이 많아 식물의 식생이나 분포 정보를 얻기에 어려움이 많으며 같은 표본이 2-3개는 기본이고 10개씩 있는 경우도 있는 등 복제표본(duplicate) 비율도 높았다. 또한 채집지가 산림과학원이었음에도 오동정된 표본이 많아 종에 대한 정보를 주는데도 적절치 않았다. 산림과학원 표본을 기증 받은 이후 데이터 입력 및 관리가 제대로 되지 않아 DB상의 큰 혼란을 가져오고 있음도 확인할 수 있었다. 국립수목원(KH) 물푸레나무속의 경우 실제 표본수인 1,158점 중에 300여점인 26%정도가 산림과학원 표본이었는데 대부분이 입력이 안 되어 있거나 입력이 되어 있어도 부분적으로 입력되어 확인이 불가능하였다. 기증을 받을 당시 DB입력이 제대로 되지 않았고, 추후에도 작업이 이루어지지 않은 것으로 추정된다.

문제는 표본이 있고 없고 또는 어디에 있느냐가 아니라 현재 표본의 상태에 대해 파악이 되지 않는다는 점이다. 표본 자료를 데이터베이스화 한다는 것은 단순히 정보만을 디지털방식으로 전환한다는 것이 아니라 모든 관리의 과정이 함께 가야 한다. 표본이 대여되었을 경우, DB는 몇 점의 표본이 언제 어느 곳으로 대여되었으며 언제 다시 돌려받을 예정인

지까지도 보여주어야 한다. 기증 받은 표본에 대해서도 소장 기준을 정해 기준에 적합한 표본은 바로 데이터베이스화 시켜야 한다. 그래야 어떤 분류군이 몇 점이며 복제표본은 어느 정도 되는지, 공간 정보 등 기본 정보가 부족한 표본들은 어떻게 보완해서 소장할 것인지에 대해 파악하고 관리할 수 있다. 이 같은 방식은 시간을 줄여줄 뿐만 아니라 여러 번 작업으로 인한 오류를 예방하는 방법이기도 하다.

2. 오류의 유형별 정제 방법에 대한 논의

오류별로 살펴보면 명명법적 오류의 경우 좀쇠물푸레나무의 경우처럼 이명처리 되어 다른 분류군으로 합쳐진 분류군이 DB상이나 폴더에 그대로 존재고 있는데 이는 이명처리 단계에서 작업에 혼란이 있어 일부 표본 정보만 DB가 변경되거나 일부 표본만 폴더를 이동한 것을 보여준다. 또한 수개회나무[*Syringa reticulate* f. *bracteata* (Nakai) T. B. Lee]와 같은 비합법명이 하나의 분류군으로 관리되고 있는 경우도 있었다. 본 연구에서는 자세히 다루지 못하였으나 이와 같은 명명법적인 오류들은 우리나라에서는 간과되고 있으나 국외에서는 명백한 오류로 점검, 수정되고 있다. 데이터베이스를 통해 철자 오류 등의 기초적인 오류는 쉽게 탐지하고 정제할 수 있으며 학명 문제도 정이명을 정리하여 입력하면 일괄적으로 정리가 가능하다.

공간적 오류는 Goodchild와 Clarke(2002)가 나눈 공간 자료 품질의 다섯 가지 측면을 통해 살펴 보고자 한다. 첫 번째 측면은 계보(lineage)로 이는 데이터세트(dataset)을 만들어 내는 과정에 대한 정보로 측정을 위한 도구, 자료 내용 정의에 대한 기준 등을 말하며 두 번째는 완성도(completeness)로 기대하는 자료 혹은 목표하는 자료의 얼마만큼을 충족하고 있는가이며 공간 정보의 구성이 충분한 정도로 볼 수 있다. 세 번째는 논리적 일관성(logical consistency)으로 자료 내에서의 일관성을 말하며 자료 내에서 서로 충돌이 되는 내용이 없는지를 살펴본다. 네 번째는 항목의 정확성(attribute accuracy)으로 오타나 비어

있는 항목이 없는지를 보며 각 항목별로 자료가 잘 들어가 있는지에 대한 정확성의 관점으로도 접근 할 수 있다. 마지막으로는 위치적 정확성(positional accuracy)은 좌표의 정보 같은 공간적 정보가 얼마나 정확하게 그 지점을 나타내는가로 예를 들어 좌표와 지역이 일치하는지 등으로 볼 수 있다.

표본관 공간자료 품질 평가 기준 중 현재 국내표본관 현황에 적용 가능한 3가지 기준으로 KH 물푸레나무속 DB를 평가해 보았을 때, 공간정보간의 일관성을 평가하는 논리적 일관성은 문제가 없었다. 즉, 행정구역 간 충돌이나 행정구역과 지명간의 불일치는 찾아볼 수 없었다. 항목의 정확성은 좌표 정보를 포함한 다른 정보(국명, 행정구역명 등)까지 모두 포함하여 평가하는데 좌표 정보가 부재한 자료가 많아 위치적 정확성과 차별화가 되지 않았다. 위치적 정확성은 각 좌표와 행정구역의 일치 여부까지도 점검해야 하나 본 연구에서는 좌표 정보를 가지면 위치적 정확성을 가지는 것으로 판단하였음에도 정확성은 48.38%로 이는 좌표 정보의 부족을 다시금 보여준다.

이러한 공간 정보의 오류는 표본의 공간 정보에 좌표를 부여하는 지리참조연산(georeferencing)을 통해 정제가 가능하다. 지리참조연산 방법은 분류학적인 오류의 정제에 비해 수월하며 자동화된 프로그램을 사용할 수도 있다. KH 물푸레나무속 데이터를 살펴본 결과, 본 연구에서 시행하지는 못했으나 지리참조연산 작업을 실시한다면 논리적 일관성과

항목의 정확성, 위치적 정확성을 각각 100%, 98.5%, 97.7%로 올려 품질을 향상시킬 수 있을 것으로 기대된다(Figure 19).

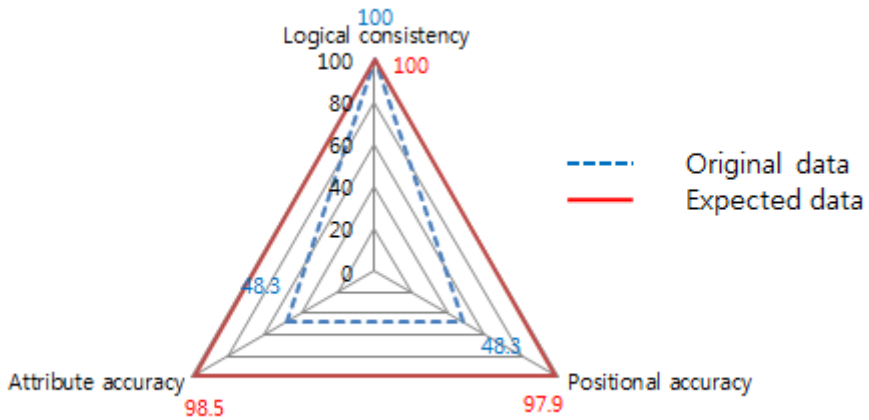


Figure 19. Quality comparison between the original dataset(dotted blue line) and the expected one(solid red line). The quality of original dataset can be improved considerably by data cleaning such as georeferencing.

분류학적인 오류의 경우 조사대상 전체의 평균 오동정률은 10.4% 정도였는데 중요한 것은 오동정이 없는 분류군부터 67.1%로 매우 높은 분류군들이 섞여 있으며 오동정률에 있어 분류군별 경향성을 찾기가 어렵다는 점이다. 이는 자료를 사용함에 있어 신뢰도를 크게 떨어뜨리며 자료 사용 자체에 대한 혼란을 야기할 수도 있다.

지역별로는 한정되거나 좁은 지역에 분포하는 종의 오동정률이 가장 높았는데 일반적으로 생각할 때 좁은 지역에 분포하면 동정이 더 쉬울 것 같지만 실제로는 유사한 다른 종들이 함께 분포하고 있기 때문에 지

역에 대한 선입견에 의한 채집으로 인해 동정의 오류가 발생하는 것으로 보았다. 흔한 식물을 덜 흔한 식물로 동정하는 비율이 높은 부분에 대해서는 우선 흔한 식물의 채집 빈도가 더 높아서 일 것이고 두 번째로는 지역별 채집에서와 유사하게 흔하지 않은 식물이 분포하는 지역에 가면 선입견으로 흔한 식물을 흔하지 않은 식물로 식별한 결과로 생각된다. 변이에 의해 개체가 평소 보던 것과 다르게 보였다면 확률은 더 높아질 것이다. 또한 흔하지 않은 분류군의 존재에 대해 알지 못한 경우도 있을 것이다. 채집을 한 후에라도 도감이나 검색표를 이용하여 확인하는 작업이 반드시 필요하다.

이런 오동정에 대한 정제 작업으로 국립수목원은 2012년과 2015년 두 번에 걸쳐 표본 재동정 사업을 실시하였다. 그 중 2015년은 수목표본을 대상으로 하였으며 본 연구의 조사 대상이었던 물푸레나무과, 피나무과와 녹나무과도 포함이 되었다. 당시 물푸레나무과 표본 중 689점이 오동정으로 평가 되었는데 역으로 계산해서 이 오동정 수를 본 연구에 포함시키면 오동정률은 8.2%에서 31.5%로 높아진다. 마찬가지로 피나무과는 0.6%에서 14.8%로 높아진다. 녹나무과는 이번 연구에서 일부만 조사하였으므로 비교하기는 어려우나 현재 오동정률이 2.6%이고 당시 오동정률은 6.3%였다. 따라서 재동정은 한번의 과제로 끝나는 것이 아니라 시스템에 의해 정기적으로 이루어져야 궁극의 신뢰성을 확보할 수 있다.

또한 분류학적 오류의 정제에 있어 정밀성의 PDE와 PTD를 이용하되 결과보다도 왜 그런 차이가 발생하였는지에 집중하여야 한다. 결과에서 살펴본 물푸레나무속과 수수꽃다리속의 경우를 가지고 가정해 보자. 만약 기관에서 숫자의 차(PDE)의 품질 측정 목표치(MQO, Measurement quality objectives)를 5%로 설정한다면 개회나무와 꽃개회나무는 기준에서 벗어나는 분류군이 되는 것이고 2015년도와 2017년도에 표본을 재동정 할 때 어떤 부분이 달랐기 때문에 이러한 표본수의 차이가 발생하였는지를 집중적으로 살펴보아야 한다(Figure 20). 마찬가지로 표본관에서 두 동정자(혹은 연구실) 간에 불일치 정도(PTD)의 기준을 15%로 잡는다고 가정하자. 들메나무(48.3%)와 물들메나무(39.2%)는 기준치보다 불일치율이 140%이상 초과된 것이므로 주의 깊게 보아야 한다(Figure 21). 중요한 것은 왜 두 분류학자 간에 차이가 발생하는지를 보는 것이므로(Stribling et al., 2008), 기준을 초과하는 들메나무와 물들메나무 표본의(Stribling et al., 2003) 어떤 부분이 동정의 불일치를 일으키는지를 파악해야 한다. 불일치를 보였던 표본들이 꽃이나 열매 없이 잎으로만 제작되어 동정의 어려움을 일으킨 것일 수도 있고, 두 동정자가 이용하는 검색표나 도감 등의 기준의 차이에 의한 것일 수도 있다. 혹은 분류학적으로 정리가 되지 않은 분류군에 대해서도 이런 불일치는 발생할 수 있다. 관리자는 이런 분석을 통해 원인을 밝히고 불일치가 높은 분류군의 경우에는 신규 채집물이 들어올 경우 동정에 좀 더 신경을 쓰는 등 향후 관리에 이용할 수 있다.

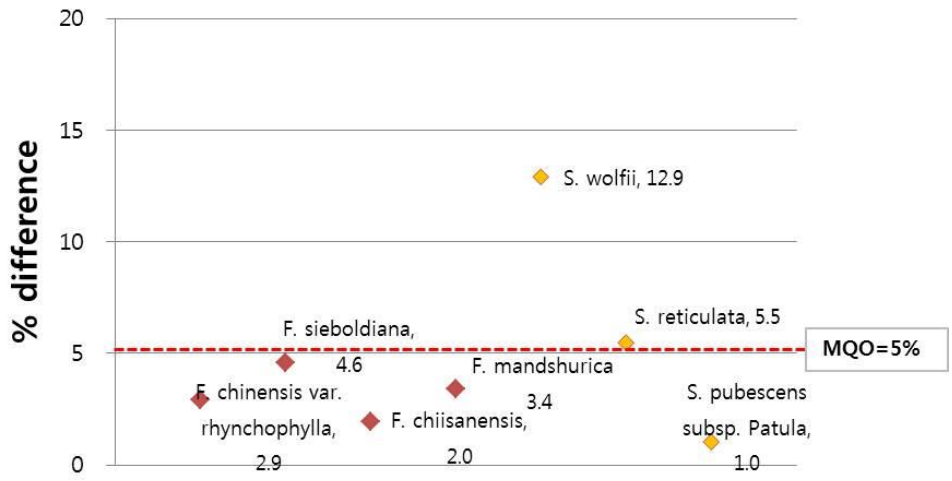


Figure 20 . Comparison of percentage difference in enumeration(PDE) & MQO.

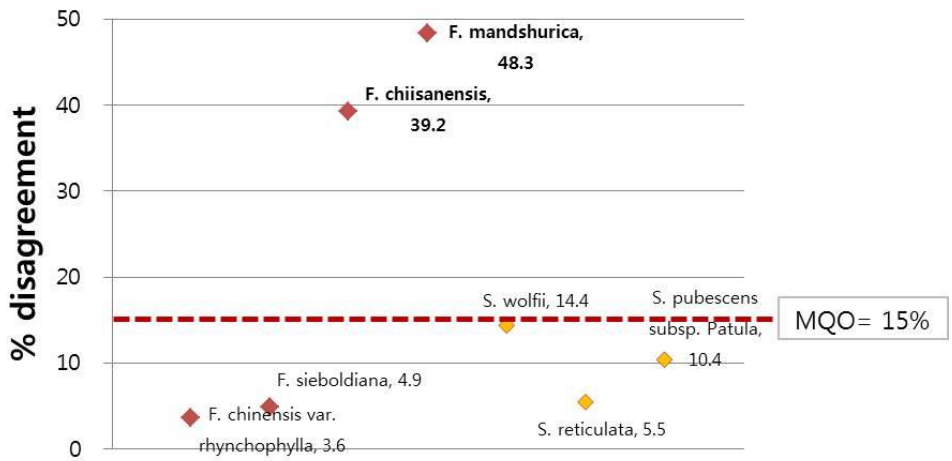


Figure 21. Comparison of percentage taxonomic disagreement(PTD) & MQO.

최근 몇 년간 국립 표본관의 해외표본 수집이 크게 늘고 있다. 그러나 국내에 전문가가 많지 않아서 동정을 하지 못하고 보관만 되어 있는 표본들이 많다. 이러한 해외 표본의 경우 특히 두 곳 이상의 연구실에 동정을 맡기고 정밀성에 대한 기준 설정과 측정을 하는 것이 반드시 필요하며 이를 통해 표본과 자료의 신뢰도를 제고할 수 있을 것이다. 또한 재동정 작업도 별도의 작업 결과를 받아 다시 기관에서 입력하는 방식이 아니라 디지털화된 DB를 재동정자에게 주고(혹은 온라인 상으로 공유하고) 재동정 작업의 결과를 바로 입력할 수 있도록 해야 한다.

분류학적 오류의 점검 과정 중 분포도 작업에서 들메나무는 가장 극단적으로 자료정제의 필요성을 보여주었다. 정제 전 분포만 본다면 들메나무는 백두대간 뿐만 아닌 경기, 충청, 전라, 경상도, 제주도까지 거의 전국에 걸쳐 분포하는 종이다. 그러나 정제 후 들메나무의 분포지는 명확하게 강원도 백두대간 지역 및 덕유산 지역임을 알 수 있다(Figure 22).

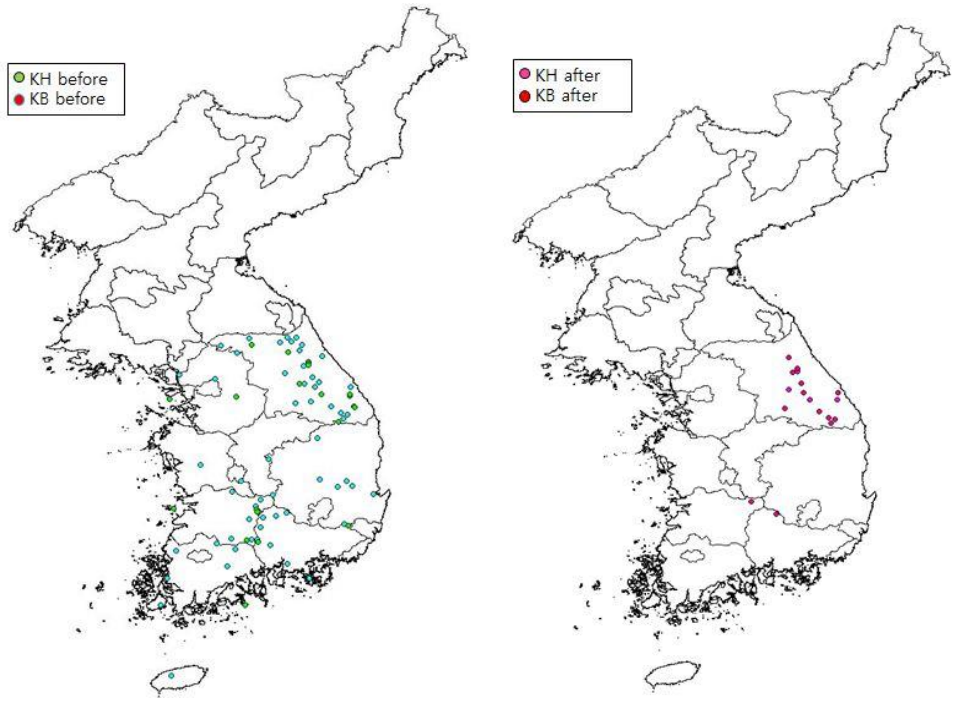


Figure 22. The effect of data cleaning. Before cleaning, *Fraxinus mandshurica* Rupr. seems to be distributed in all of Korean peninsula(left) but the actual distribution area is limited to Beakdudaegan(right).

분류학적인 오류는 가장 까다로우며 아직까지 전문가가 직접 동정을 하는 것 이외의 뚜렷한 정제 방법에 대한 논의가 없었다. 그러나 본 연구의 결과는 신뢰할 수 있는 기존 자료의 분포지 정보를 통해 부분적인 오동정 탐지가 가능함을 보여주고 있다.

분포가 의심스러운 표본이 있는 종들만을 대상으로 결과를 다시 살펴보면 오동정 예측에 의미가 있었다. 오동정으로 의심했던 대상 중 실제 오동정인 비율은 24종 중 13종이 100%, 1종이 98.5%로 100%에 가까운 예측률을 보였다(Figure 23). 이는 분포를 통한 오동정의

예측이 전 종을 대상으로 하였을 때는 큰 의미가 없는 것처럼 보이나 어느 정도 일정한 분포지역을 가지고 있을 종에 대해서는 오동정의 상당 부분을 미리 예측하고 정제할 수 있음을 보여준다.

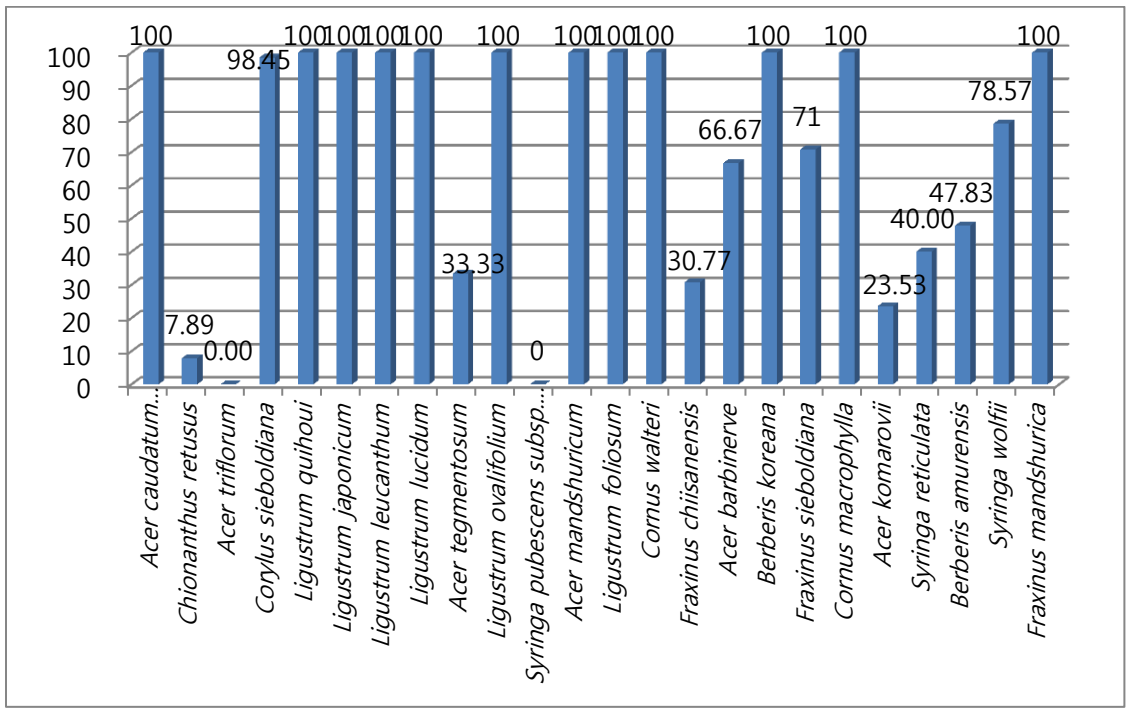


Figure 23. The rate of detecting misidentification. Thirteen out of twenty four species present 100% of detecting misidentification rate.

3. 관리 절차 단계별 논의

Chapman(2005)이 제시한 자료의 관리 절차의 단계(Table 7)별로 국내 표본관 자료 관리의 문제점과 해결방안을 논의해보고자 한다.

표본을 수집하는 시점에서의 데이터 기록 단계에서는 식물에 대한 오동정과 위치 정보에 대한 기록의 누락이 있었다. 표본 채집 전 채집자에게 라벨에 기재되어야 할 사항을 정확하게 알려주고 채집 시 필요한 정보를 모두 기재하도록 해야 한다. 디지털화 이전 데이터 수작업단계에서는 라벨 제작 시 자료의 누락이나 오타의 발생할 수 있는데 이러한 오류는 미미하였다.

채집해 온 표본은 각 표본관의 확정 시스템(determination system)에 의해 채집 시 기록된 동정에 대해 2명 이상의 전문가가 한 번 더 확인하고 최종적인 동정의 결과를 기록하게 되어 있다. 그러나 이 과정을 거쳤음에도 불구하고, 또는 이 과정이 누락됨으로써 채집물의 동정과 기록의 단계에서 분류학적 오류가 주로 발생하고 있다. 현재 확정 시스템이 잘 운영되고 있는지, 잘 운영되고 있음에도 문제가 발생한다면 그 원인이 무엇인지 내부적인 점검을 통한 원인 파악과 수정이 필요하다. 또한 누락된 정보가 있는지 확인하고 행정구역 등의 위치정보의 기재가 정확한지, 불일치는 없는지 확인해야 한다.

전문가에 의한 확정(determination)작업이 원칙이나 일부 표본은 제작이나 기증과 동시에 라벨작업만 거쳐 바로 수장되고 있다. 이는 관

리의 문제임의 동시에 표본관의 예산과 전문인력의 부족이 원인으로 바로 해결되기 어려운 것이 현실이다. 이에 근본적인 해결방안은 아니더라도 현재의 인력을 활용하되 오류를 줄일 수 현실적인 절차를 제안한다 (Figure 24). 이 절차를 따른다면 분류학적 지식이 없더라도 분포로 탐지할 수 있는 식별의 오류와 좌표정보의 오류를 데이터의 입력과 동시에 정제할 수 있을 것으로 기대한다.

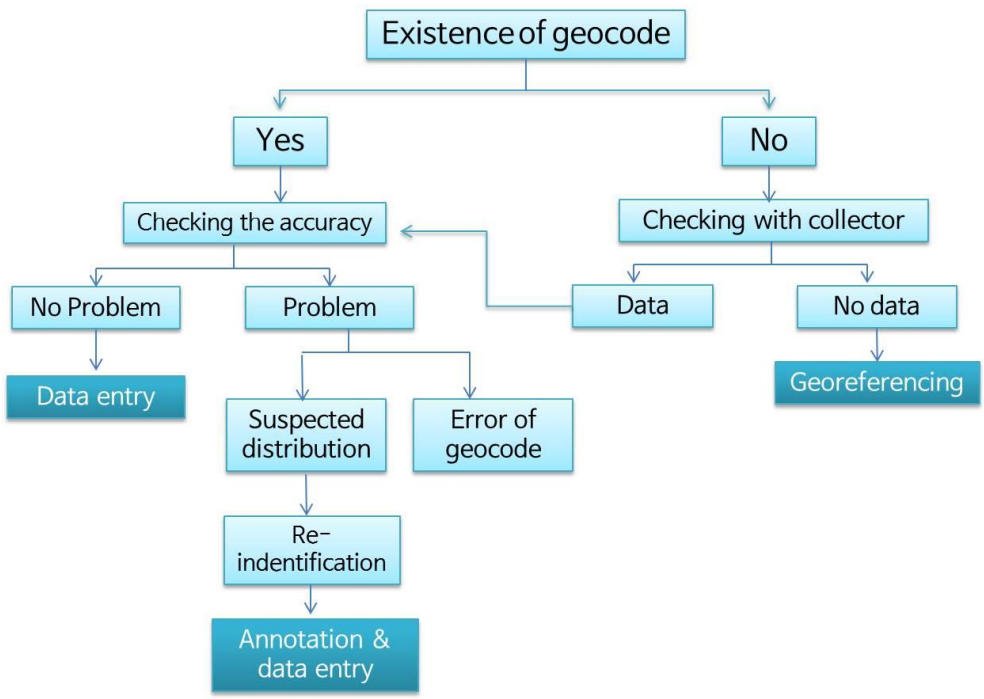


Figure 24. The process for data entry of spatial data. Through this process even non-professional workers can clean spatial and taxonomic errors.

자료의 디지털화 단계에서는 자료가 누락되거나 명백하게 문제가 있는 자료가 입력이 되거나 입력의 실수도 발생하고 있었다. 국립수목원 소장 물푸레나무속의 경우 물푸레나무는 6개의 표본, 쇠물푸레나무의 경우 4개의 표본이 바코드가 중복되었다. 일차적으로는 자료의 입력이 주로 계약직 직원 혹은 일용직에 의해 이루어지다 보니 책임감 혹은 전문성의 부족 등이 원인으로 작용하였다고 본다. 보다 근본적으로는 바코드의 인쇄 및 부착, 자료 입력 후 확인을 거치는 시스템이 없거나 자리 잡지 못한 것이 원인이라고 본다.

마지막 단계는 축적한 자료를 공개하고 이용하는 단계로 현재 국내 표본관은 표본 정보의 DB를 구축하고 온라인상에 공개만 하였을 뿐 적극적인 활용은 하지 못하고 있다. DB의 구축보다 중요한 것은 체계적인 관리와 활용이다. 디지털화하여 시스템을 구축하였다고 한들 실제로 이용하지 않고 쉽게 이용할 수 없다면 죽은 정보와 다를 바가 없다. 정보의 축적과 디지털화와 더불어 표본의 큐레이션 방향과 방법도 이에 따라 변화해야만 한다(Scoble and Bourgoin, 2010). 사용자와의 소통을 통한 오류의 탐지와 정제의 필요성에 대해서는 꾸준히 이야기되어 온 만큼 (Orr 1998,Stribling et al., 2003) 온라인 상에서 전문가와 비전문가 과 사이에 끊임 없는 소통을 통해 DB가 지속적으로 업데이트 되어야 한다. 자료 정제를 통해 DB와 실제 표본과의 간격을 줄여 신뢰도를 높이며 누구나 온라인에서 정보에 쉽게 접근하고 다운받을 수 있도록 해야 한다. 기존의 식물표본 관리와 DB구축에서 한 발 더 나아가 DB를 국내뿐 아

니라 GIBF와 같은 국제적 사이트에 배포하고 운영, 발전시켜나가는 큐레이션이 이루어져야 할 것이다.

Table 7. The stages of data management process(Chapman, 2005).

▪ Data capture and recording at the time of gathering
▪ Data manipulation prior to digitization
▪ Identification of the collection and its recording
▪ Digitization of the data
▪ Documentation of the data
▪ Data storage and archiving
▪ Data presentation and dissemination
▪ Using the data

위의 단계에 오류의 점검과 정제 및 정제한 내용에 대한 문서화, 그것을 추후 관리의 단계가 추가하고자 한다. 자료의 정제 과정은 오류의 원인을 밝히는 차원에서 중요하며 그 결과를 통해 같은 오류가 다시 발생하지 않는 절차로 가야 하고 같은 맥락에서 오류의 정제와 오류의 예방은 반드시 같은 시점에 진행되어야 한다(Chapman, 2005). 관리자는 이런 오류 탐지와 정제 과정, 그 결과의 기록, 분석을 통해 관리 지침을 수정하거나 추가하는 등 향후 오류 예방을 위한 대책을 바로 마련하고 시스템에 적용시켜야 한다. 모든 과정은 데이터베이스를 기반으로 진행되고 기록되어야 한다.

마지막으로 현재 국립 표본관은 모두 전담 큐레이터 없이 연구사가 담당하고 있다. 표본관은 죽어 있는 공간이며 단순한 관리의 대상이 아니라 수백 년의 자료의 보고이다. 이 자료를 살아 움직이게 하고 널리 이용될 수 있도록 하려면 표본관의 모든 과정을 오롯이 관리할 수 있는 큐레이터가 꼭 필요하다.

제 6 장 결 론

국립표본관 식물표본과 DB를 살펴본 결과, 명명법적인 오류에 있어 이명이나 비합법명이 사용되고 있었다. 공간적인 오류에서는 국립수목원(KH)는 채집 년도가 없는 자료가 17%, 좌표 정보가 없는 자료는 47%였으며 국립생물자원관(KB)의 경우는 채집 년도는 거의 누락이 없었고 좌표정보는 4% 정도 없었다. 분류학적 오류에 있어서 오동정률은 대략 5%에서 25% 정도의 분포를 보였으며 평균은 10.4%였다. 그러나 오동정이 없는 분류군부터 67.07%인 분류군까지 편차가 컸으며 이러한 편차는 자료의 신뢰도를 떨어뜨리고 있었다.

분류학적 오류의 경우 한 속 내에 종수가 3종일 경우 오동정이 가장 많았으며 한정된 지역에 분포하거나 백두대간에 분포하는 분류군의 오동정률이 높았으며 정확성에 있어서는 91.5%의 오동정이 같은 속 내에서 다른 종으로 동정된 것이었다. 분류학적 오류는 유사한 형태적 특징, 채집 지역에 선입견이나 종에 대한 지식의 부족, 채집 빈도의 차이에 의한 것이었다. 이런 분류학적 오류는 신뢰할 수 있는 기존 분포 정보를 활용한 분포도 작업을 통해 25.6% 정도를 탐지하는 것이 가능했다.

DB관리에 있어서의 핵심은 자료의 데이터베이스화이다. 표본의 채집부터 디지털화, 오류의 점검, 정제 예방까지의 모든 관리 과정은 데이터베이스를 통해 이루어져야 하며 이를 담당하는 전담 큐레이터의 존재가 필수적이다.

인용 문헌

Aikio, S., Duncan, R.P. and Hulme, P.E. 2010. Herbarium records identify the role of long-distance spread in the spatial distribution of alien plants in New Zealand. Journal of Biogeography 37 (9): 1740-1751.

Anderson, R.P. 2012. Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. Annals of the New York Academy of Sciences 1260 (1): 66-80.

Applequist, W.L., Mcglinn, D.J., Miller, M., Long, Q.G. and Miller, J.S. 2007. How well do herbarium data predict the location of present populations? A test using Echinacea species in Missouri. Biodiversity and Conservation 16 (5): 1397-1407.

Armstrong, J. 1992. The funding base for Australian biological collections. Australian Biologist 5 (1): 80-88.

Bebber, D.P., Carine, M.A., Wood, J.R., Wortley, A.H., Harris, D.J., Prance, G.T. et al. 2010. Herbaria are a major frontier for species discovery. Proceedings of the National Academy of Sciences 107 (51): 22169-22171.

Bolmgren, K. and Lonnberg, K. 2005. Herbarium data reveal an association between fleshy fruit type and earlier flowering time. International Journal of Plant Sciences 166 (4): 663-670.

Booth, T.H., Jovanovic, T., Old, K.M. and Dudzinski, M.J. 2000. Climatic mapping to identify high-risk areas for Cylindrocladium quinqueseptatum leaf blight on eucalypts in mainland South East Asia and around the world. Environmental Pollution 108 (3): 365-372.

Brooks, R., Lee, J., Reeves, R.D. and Jaffré, T. 1977. Detection of nickeliferous rocks by analysis of herbarium specimens of indicator plants. Journal of Geochemical Exploration 7: 49-57.

Burrough, P.A. and McDonnel, R.A. 1998. Principles of Geographical Information Systems, 265 p.

Chapman, A.D. 1999. Quality control and validation of point-sourced environmental resource data. In Spatial accuracy assessment: Land information uncertainty in natural resources. K. Lowell and A. Jaton (eds.), Ann Arbor Press, Chelsea.

Chapman, A.D. 2005. Principles and methods of data cleaning. GBIF.

Chapman, A.D. 2005. Principles of data quality. GBIF.

Chavan, V. and Krishnan, S. 2003. Natural history collections: A call for national information infrastructure. CURRENT SCIENCE-BANGALORE- 84 (1): 34-42.

Chang, C.S., Jeon, J.I. and Min, W.K. 1999. The distribution of the woody plants of South Korea based on herbarium material of Kwanak Arboretum (V) -Oleaceae. Bulletin of the Seoul National University Arboretum 19: 1-28

Chang, C.S., Chang, K.S., Ahn, Y.S. and Kim, H. 2012. 1 Artiele : Georeferencing of Primary Species Occurrence Data and Necessity of Data Quality Control -A Case Study of Two Varieties of Ox-Knee, Achyranthes bidentata Blume- Journal of Korean Forest Society 101(2): 185-194.

Chung, G.Y. and 12. 2015. Re-indentification of woody plants specimen in Korea National Herbarium.

Costa, H., Foody, G.M., Jiménez, S. and Silva, L. 2015. Impacts of species misidentification on species distribution modeling with presence-only data. ISPRS International Journal of Geo-Information 4 (4): 2496-2518.

Crawford, P.H.C. and Hoagland, B.W. 2009. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? Journal of Biogeography 36 (4): 651–661.

Crisp, M.D., Laffan, S., Linder, H.P. and Monro, A. 2001. Endemism in the Australian flora. Journal of Biogeography 28 (2): 183–198.

Davis, C.C., Willis, C.G., Connolly, B., Kelly, C. and Ellison, A.M. 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. American Journal of Botany 102 (10): 1599–1609.

Delisle, F., Lavoie, C., Jean, M. and Lachance, D. 2003. Reconstructing the spread of invasive plants: taking into account biases associated with herbarium specimens. Journal of Biogeography 30 (7): 1033–1042.

Faith, D.P., Margules, C.R., Walker, P., Stein, J. and Natera, G. 2001. Practical application of biodiversity surrogates and percentage targets for conservation in Papua New Guinea. Pacific Conservation Biology 6 (4): 289–303.

Fuentes, N., Pauchard, A., Sánchez, P., Esquivel, J. and Marticorena, A. 2013. A new comprehensive database of alien plant species in

Chile based on herbarium records. Biological Invasions 15 (4) : 847-858.

Funk, V. 2003. 100 uses for an herbarium (well at least 72). American Society of Plant Taxonomists Newsletter 17: 17-19.

Funk, V. and Richardson, K. 2002. Systematic data in biodiversity studies: use it or lose it. Systematic Biology : 303-316.

Funk, V.A., Zermoglio, M.F. and Nasir, N. 1999. Testing the use of specimen collection data and GIS in biodiversity exploration and conservation decision making in Guyana. Biodiversity & Conservation 8 (6): 727-751.

Gallagher, R., Hughes, L. and Leishman, M. 2009. Phenological trends among Australian alpine species: using herbarium records to identify climate-change indicators. Australian Journal of Botany, 57 (1): 1-9.

Goodchild, M.F. and Clarke, K.C. 2002. Data quality in massive data sets. Chapter 18: 643-659.

Goodwin, Z.A., Harris, D.J., Filer, D., Wood, J.R. and Scotland, R.W. 2015. Widespread mistaken identity in tropical plant collections. Current Biology 25 (22): R1066-R1067.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. and Peterson, A.T. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19 (9): 497-503.

Guerin, G.R. 2013 The value of herbaria to diverse collections-based research. Australasian Systematic Botany Society Newsletter 157: 43-44.

Guerin, G.R., Wen, H. and Lowe, A.J. 2012. Leaf morphology shift linked to climate change. Biology letters 8 (5): 882-886.

Hart, R., Salick, J., Ranjitkar, S. and Xu, J. 2014. Herbarium specimens show contrasting phenological responses to Himalayan climate. Proceedings of the National Academy of Sciences 111 (29): 10615-10619.

Harvard University Herbaria. 2017. https://huh.harvard.edu.

Hortal, J., Lobo, J.M. and JIMÉNEZ-VALVERDE, A. 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. Conservation biology 21 (3): 853-863.

Jiménez-Valverde, A., Lira-Noriega, A., Peterson, A.T. and Soberón, J. 2010. Marshalling existing biodiversity data to evaluate

biodiversity status and trends in planning exercises. Ecological Research 25 (5): 947–957.

Kier, G. and Barthlott, W. 2001. Measuring and mapping endemism and species richness: a new methodological approach and its application on the flora of Africa. Biodiversity & Conservation 10 (9): 1513–1529.

Kew Herbarium Catalogue. 2017.
http://apps.kew.org/herbcat/gotoHomePage.do.

Kim, H. 2014. Herbarium management and collection policy study.

Korea Biodiversity Information System. 2017.
http://www.nature.go.kr/index.jsp.

Korea National Arboretum. 2017.
http://www.forest.go.kr/newkfsweb/kfs/idx/SubIndex.do?orgId=kna&mn=KFS_15.

Korean Natural History Research Information System. 2017.
http://www.naris.go.kr/naris/main.do.

Korean National institute of biological resources. 2017.
https://www.nibr.go.kr/main/main.jsp.
https://species.nibr.go.kr/index.do.

Lee, J.W., Park, H.S. and Ahn, S.S. 2006. Korean version of Principles and methods of data cleaning. KISTI (Korea Institute of Science and Technology Information)

MacGillivray, F., Hudson, I.L. and Lowe, A.J. 2010. Herbarium collections and photographic images: alternative data sources for phenological research. Springer.

Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond Integrity Analysis. Citeseer: 200-209.

Margules, C.R. and Pressey, R.L. 2000. Systematic conservation planning. Nature 405 (6783): 243-253.

McGraw, J.B. 2001. Evidence for decline in stature of American ginseng plants from herbarium specimens. Biological Conservation 98 (1): 25-32.

Miller-Rushing, A.J., Primack, R.B., Primack, D. and Mukunda, S. 2006. Photographs and herbarium specimens as tools to document phenological changes in response to global warming. American Journal of Botany 93 (11): 1667-1674.

Murphey, P.C., Guralnick, R.P., Glaubitz, R., Neufeld, D. and Ryan, J.A. 2004. Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed

by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi). Phyloinformatics 1 (3): 1-29.

Orr, K. 1998. Data quality and systems theory. Communications of the ACM 41 (2): 66-71.

Peterson, A.T., NAVARRO-SIGÜENZA, A.G. and BENÍTEZ-DÍAZ, H. 1998. The need for continued scientific collecting; a geographic analysis of Mexican bird specimens. Ibis 140 (2): 288-294.

Redman, T.C. 1998. The impact of poor data quality on the typical enterprise. Communications of the ACM 41 (2): 79-82.

Redman, T.C. 2001. Data quality: the field guide. Digital press.

Schulman, L., Toivonen, T. and Ruokolainen, K. 2007 Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. Journal of Biogeography 34 (8): 1388-1399.

Scoble, M. and Bourgoin, T. 2010. Natural history collections digitization: rationale and value. Biodiversity Informatics 7 (2): 77-80.

Scott, W.A. and Hallam, C.J. 2003. Assessing species misidentification rates through quality assurance of vegetation monitoring. Plant Ecology 165 (1): 101-115.

Sodré, F., Fernandes, R., Moraes, M., Pougy, N., Caram, J., Dalcin, E. et al. Spatial data quality of herbarium datasets and implications for decision-making on biodiversity conservation in Brazil.

Stribling, J.B., Moulton, S.R. and Lester, G.T. 2003. Determining the quality of taxonomic data. Journal of the North American Benthological Society 22 (4): 621-631.

Stribling, J.B., Pavlik, K.L., Holdsworth, S.M. and Leppo, E.W. 2008. Data quality, performance, and uncertainty in taxonomic identification for biological assessments. Journal of the North American Benthological Society 27 (4): 906-919.

The Australasian Virtual Herbarium(AVH). http://avh.chah.org.au.

The woody plants of Korea. 2017. http://florakorea.myspecies.info/en.

Łuczaj, Ł.J. 2010. Plant identification credibility in ethnobotany: a closer look at Polish ethnographic studies. Journal of ethnobiology and ethnomedicine 6 (1): 36.

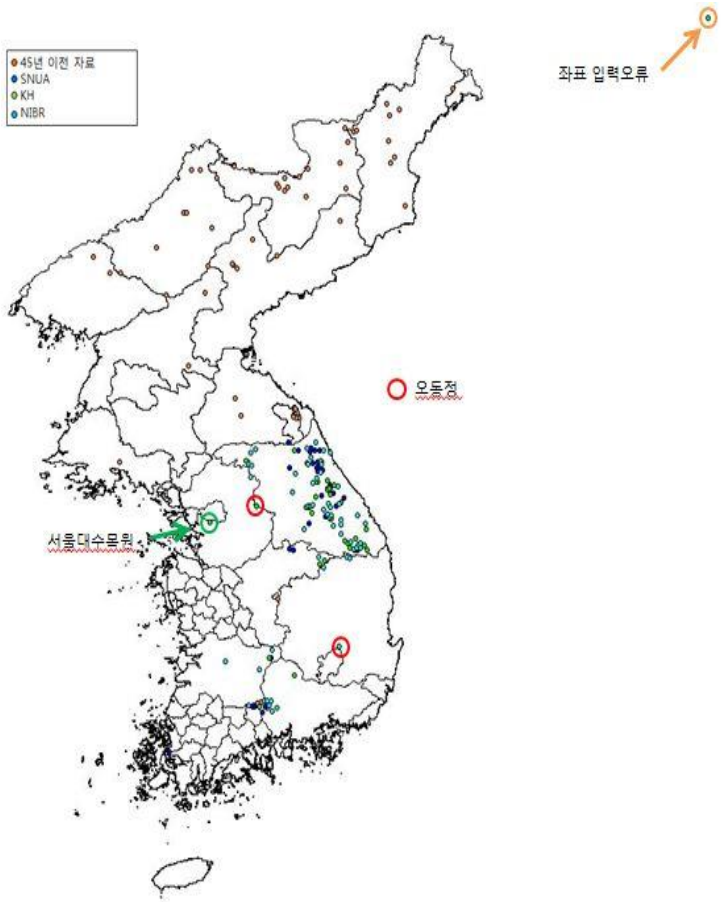
Wen, J., Ickert-Bond, S.M., Appelhans, M.S., Dorr, L.J. and Funk, V.A. 2015. Collections-based systematics: Opportunities and outlook for 2050. Journal of Systematics and Evolution 53 (6): 477-488.

Wieczorek, J., Guo, Q. and Hijmans, R. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18 (8): 745-767.

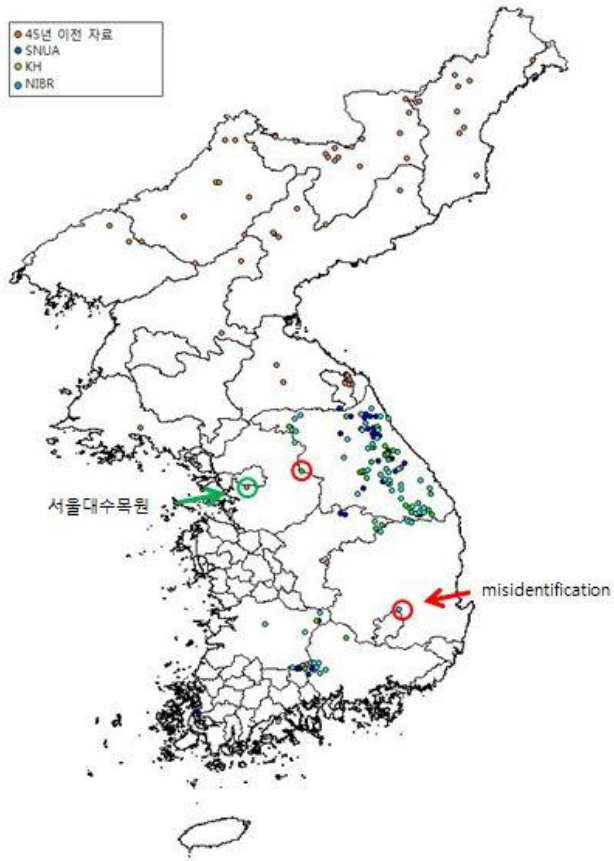
Whittaker, R.J., Araújo, M.B., Jepson, P., Ladle, R.J., Watson, J.E. and Willis, K.J. 200.5 Conservation biogeography: assessment and prospect. Diversity and distributions 11 (1): 3-23.

Williams, P., Margules, C.R. and Hilbert, D.W. 2002. Data requirements and data sources for biodiversity priority area selection. Journal of biosciences 27 (4): 327-338.

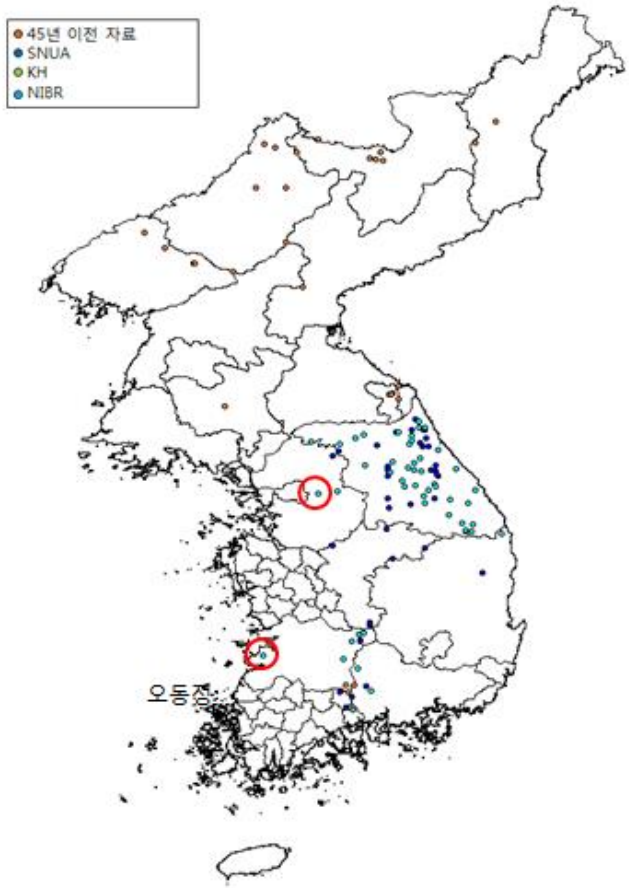
Appendix. Distribution maps of five groups.



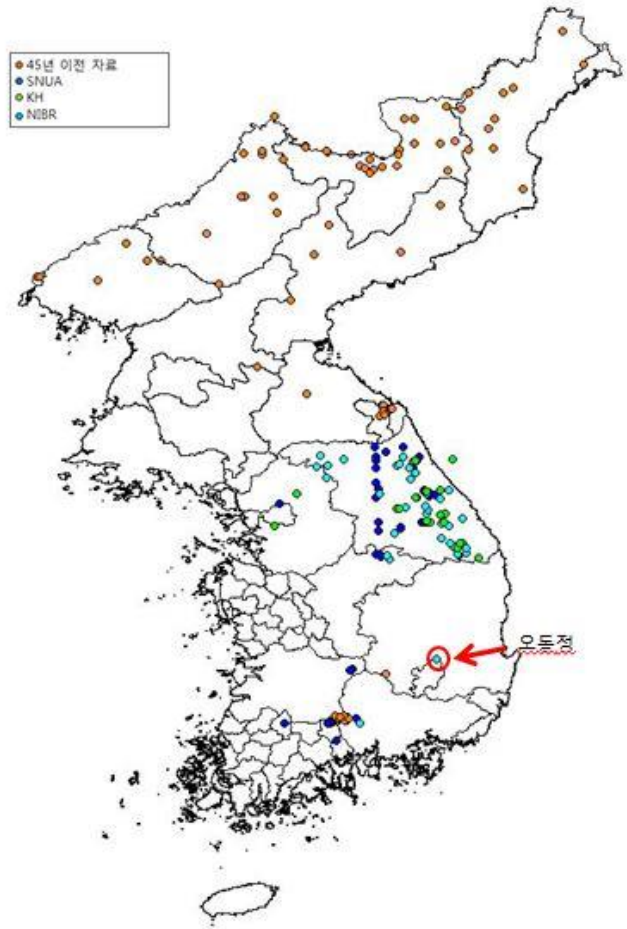
Group1. *Acer barbinerve* Maxim.



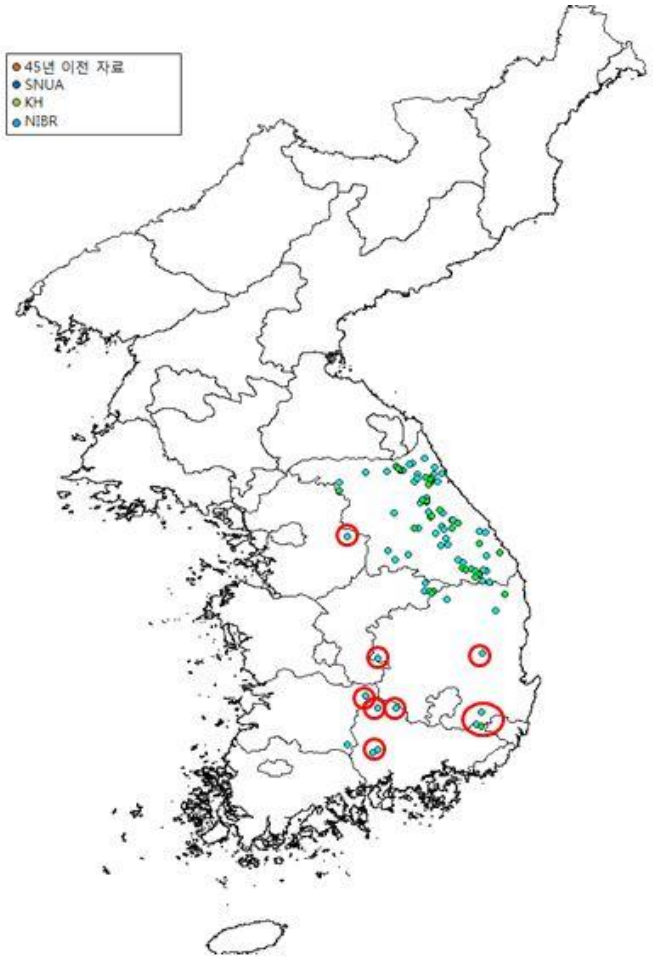
Group1. *Acer komarovii* Pojark.



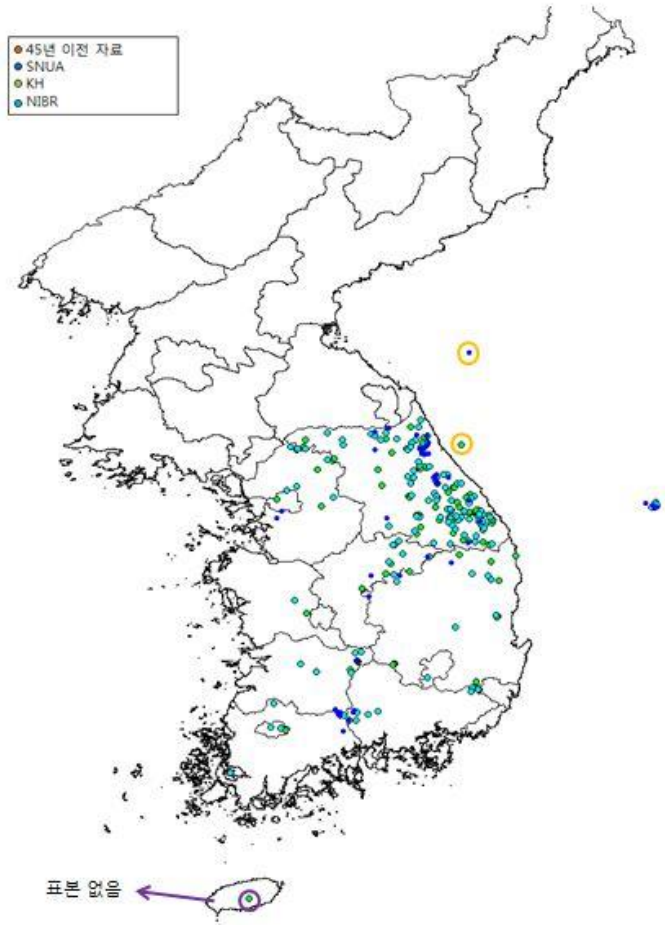
Group1. *Acer mandshuricum* Maxim.



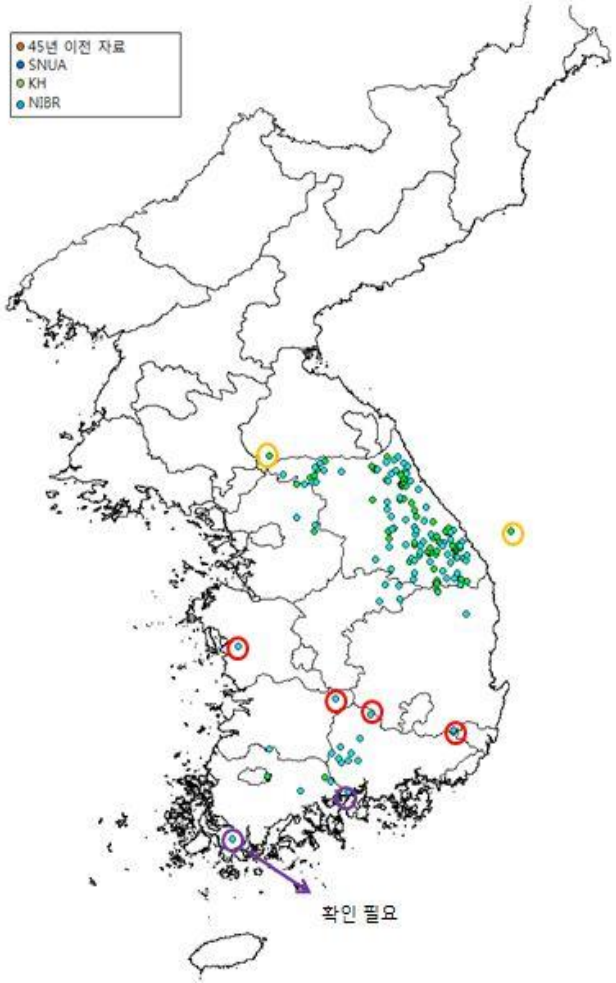
Group1. *Acer tegmentosum* Maxim.



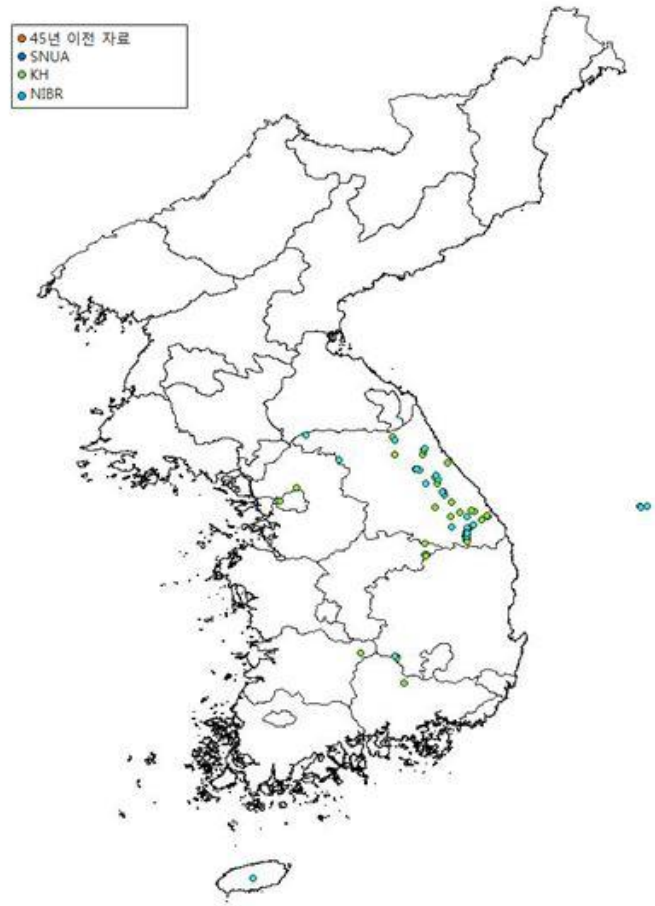
Group1. *Syringa wolfii* C,K,Schneid.



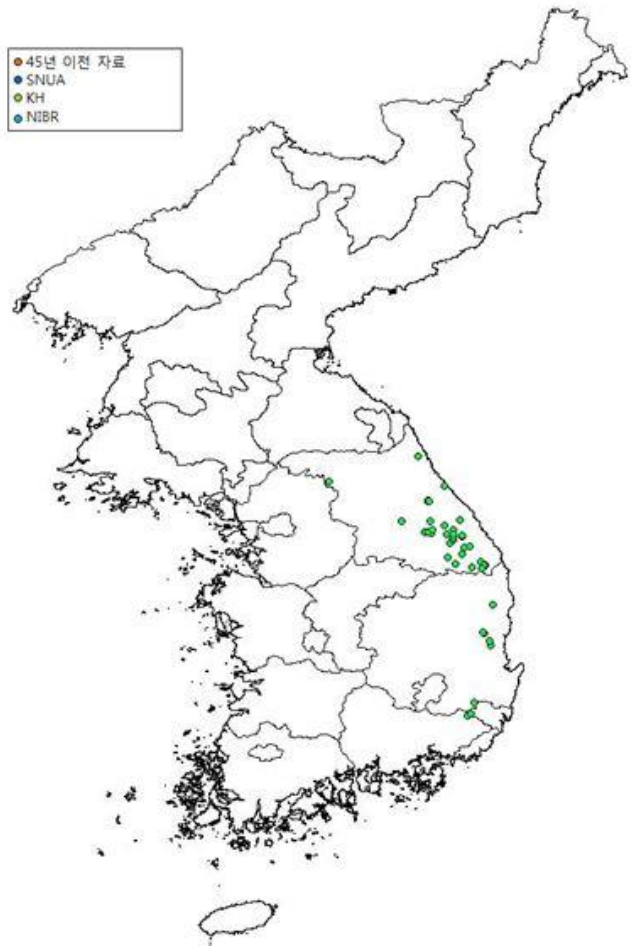
Group1. *Syringa pubecense* subsp. *patula* (Palibin) M. C. Chang & X. L. Chen



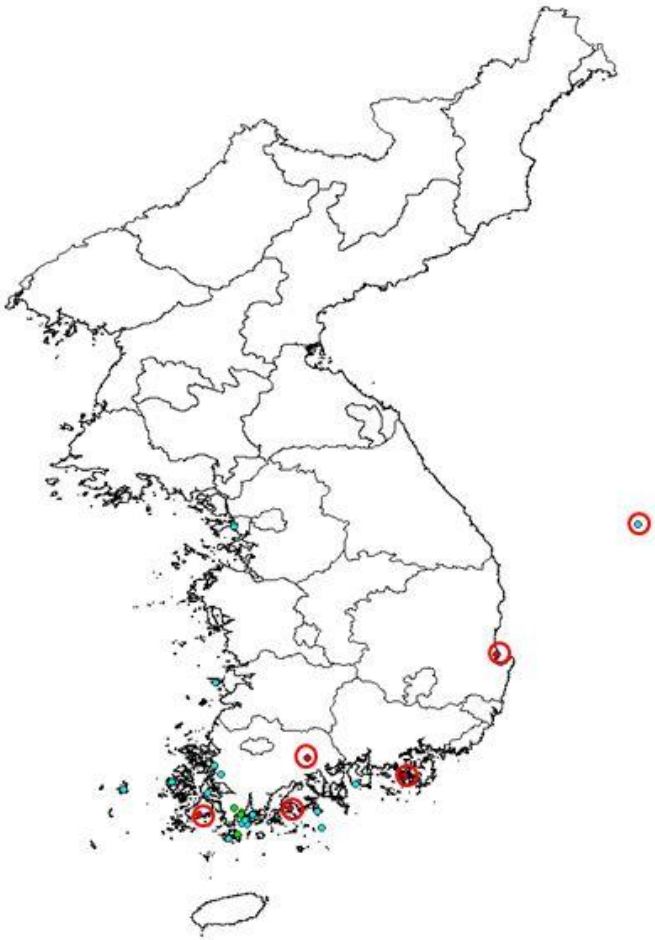
Group1. *Syringa reticulate* (Blume) H.Hara



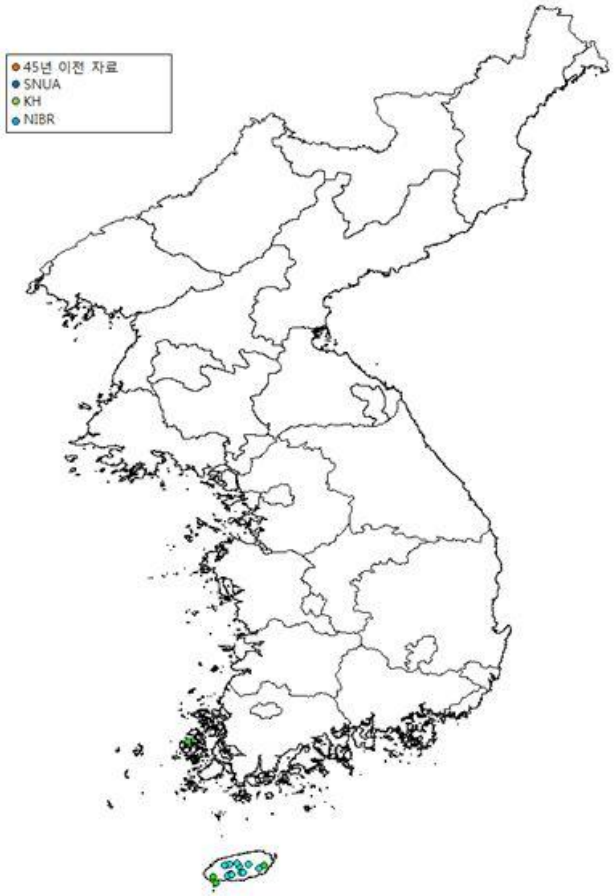
Group1. *Berberis amurensis* Rupr.



Group1. *Aristolochia mandshuriensis* Kom.

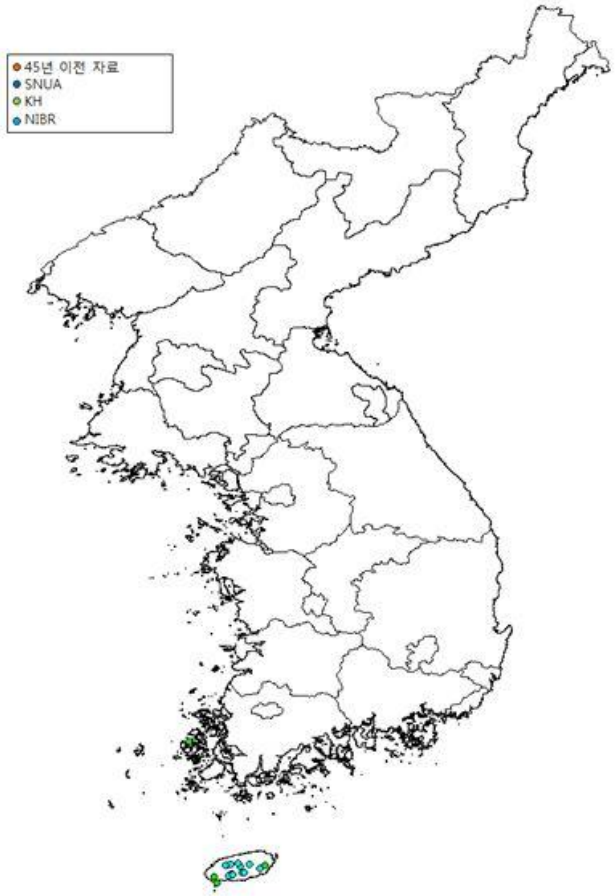


Group2. *Ligustrum quihoui* Carrière



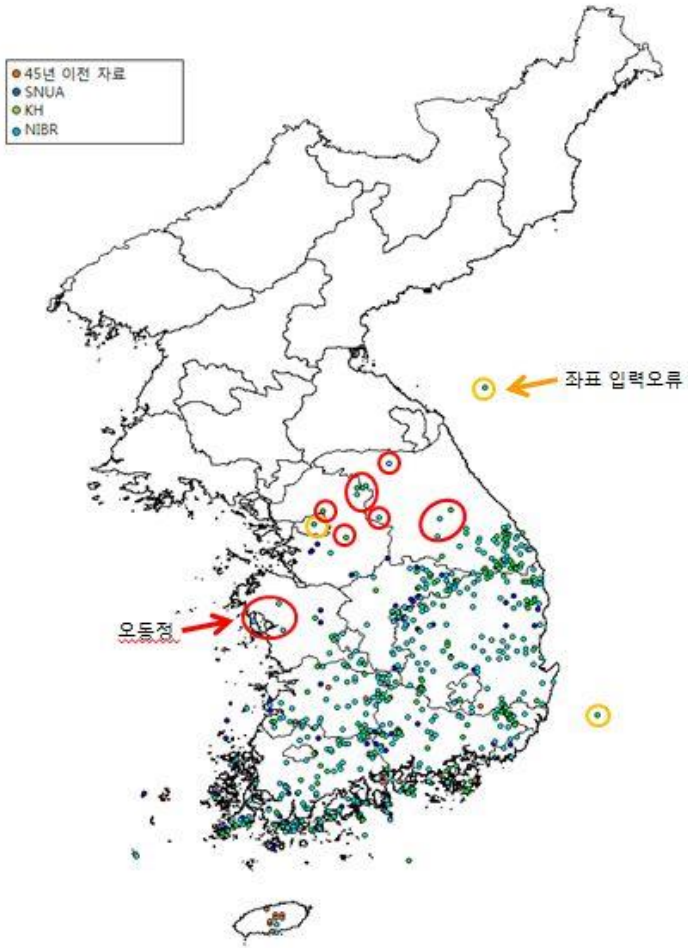
Group2. *Ligustrum obtusifolium* Siebold & Zucc.

subsp. *microphyllum* (Nakai) P.S.Green

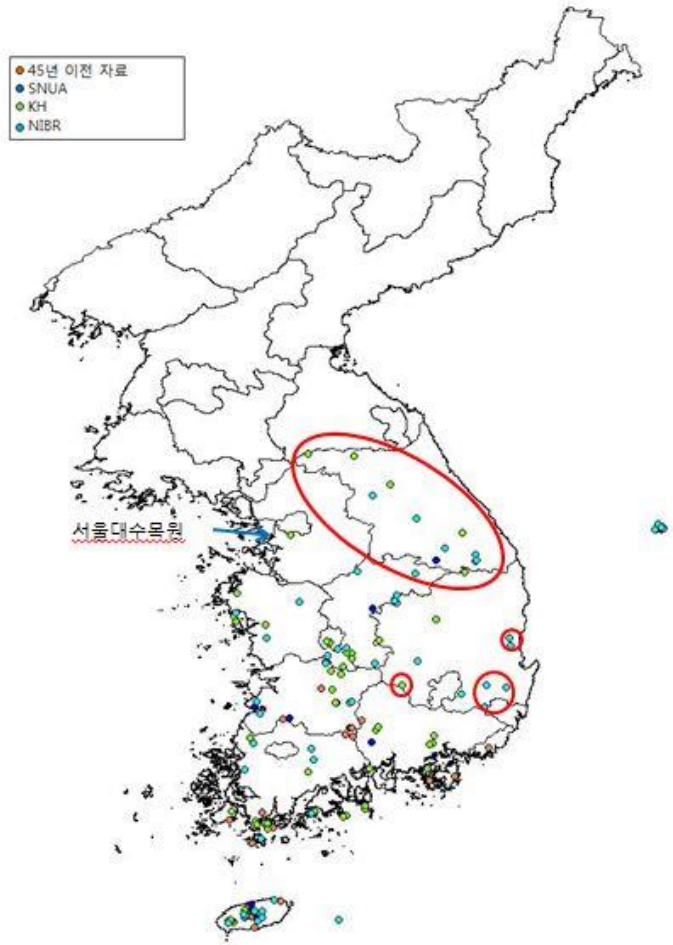


Group2.. *Ligustrum obtusifolium* Siebold & Zucc.

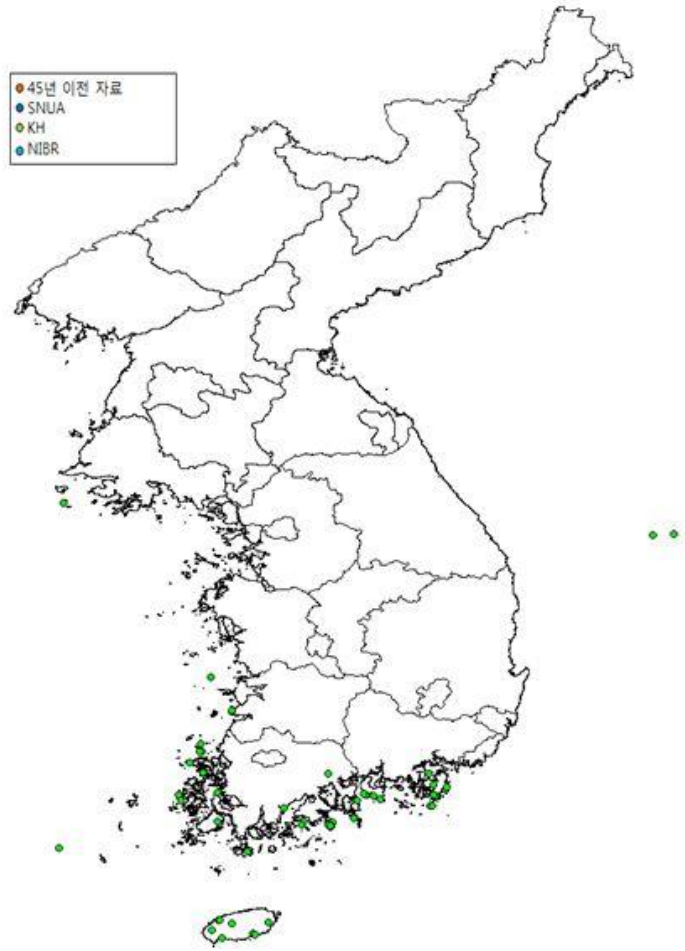
subsp. *microphyllum* (Nakai) P.S.Green



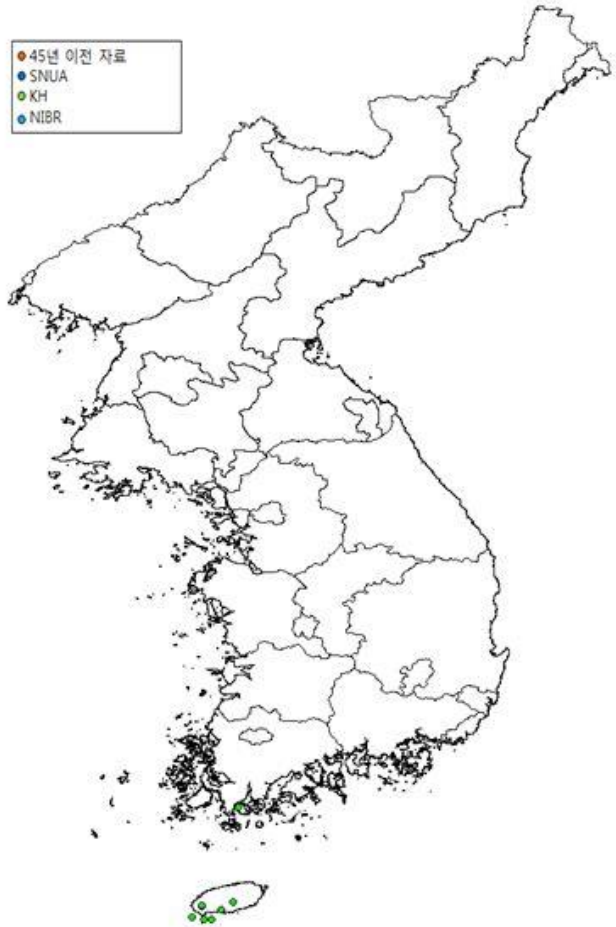
Group2. *Fraxinus sieboldiana* Blume



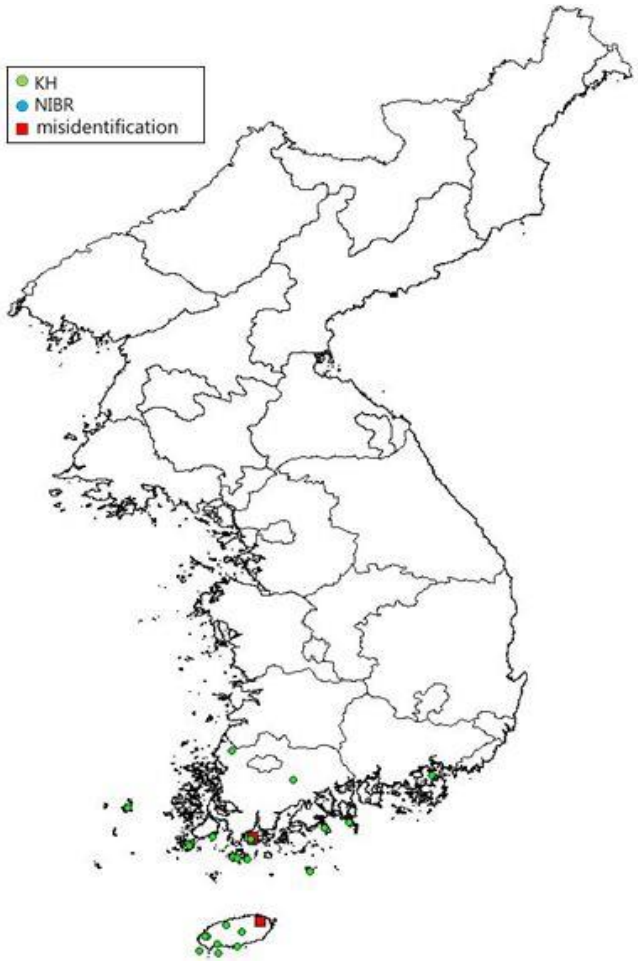
Group2. *Cornus macrophylla* Wall. .



Group2. *Machilus thunbergii* Siebold & Zucc.ex Meisn.



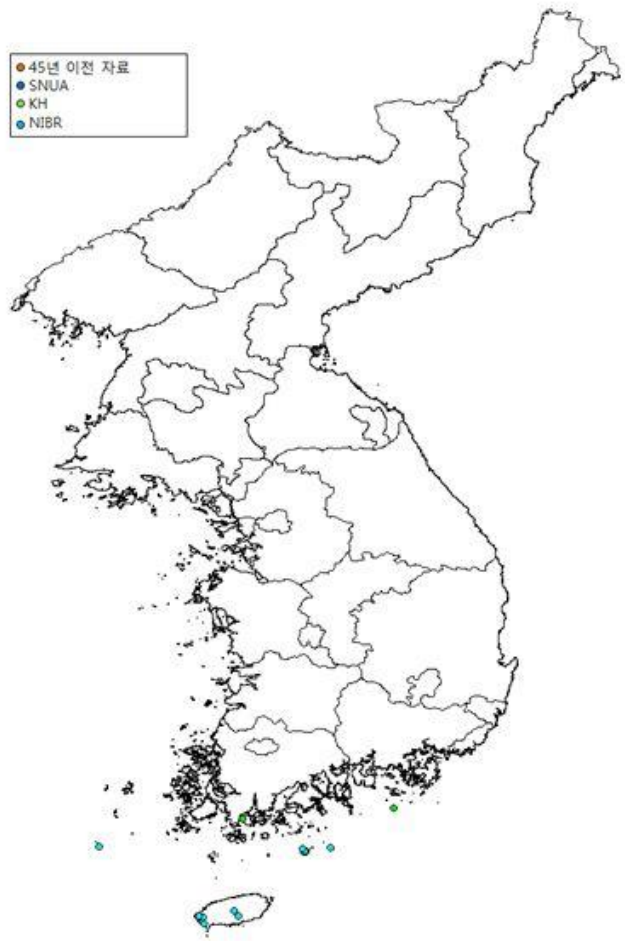
Group2. *Cinnamomum camphora* (L.)L.Presl



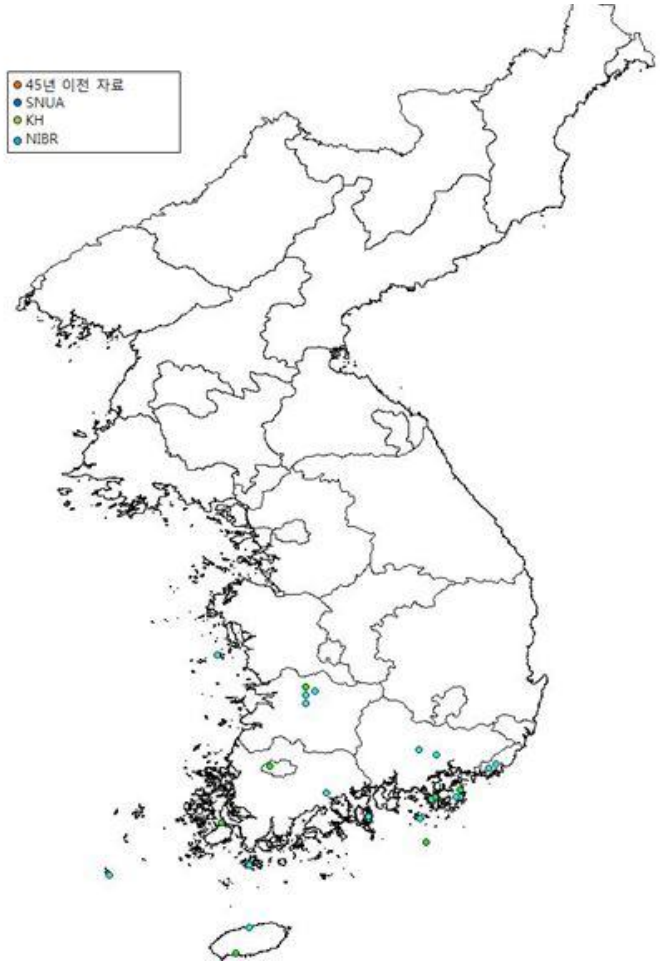
Group2. *Cinnamomum yabunikkei* H.Ohba

Group2. *Litsea coreana* H. Lév.

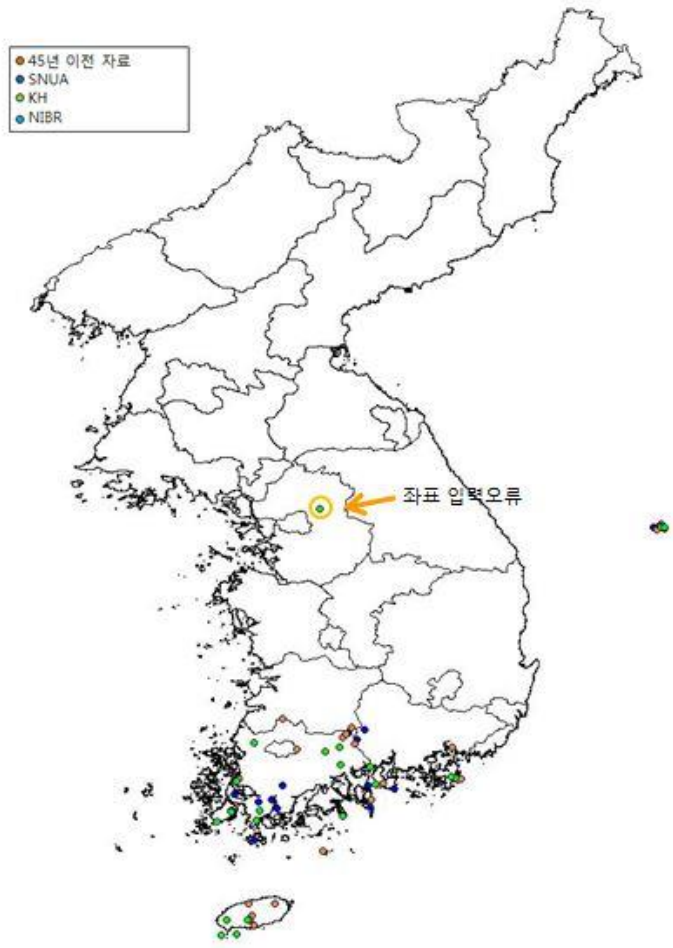
Group2. *Litsea japonica* (Thunb.) Jussieu



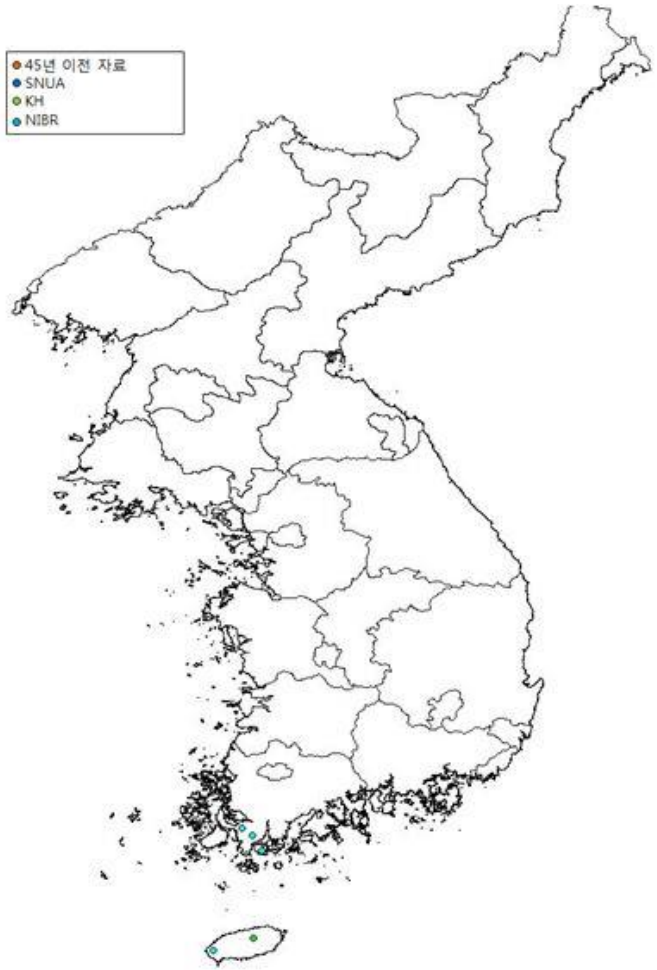
Group2. *Osmanthus insularis* Koidz.



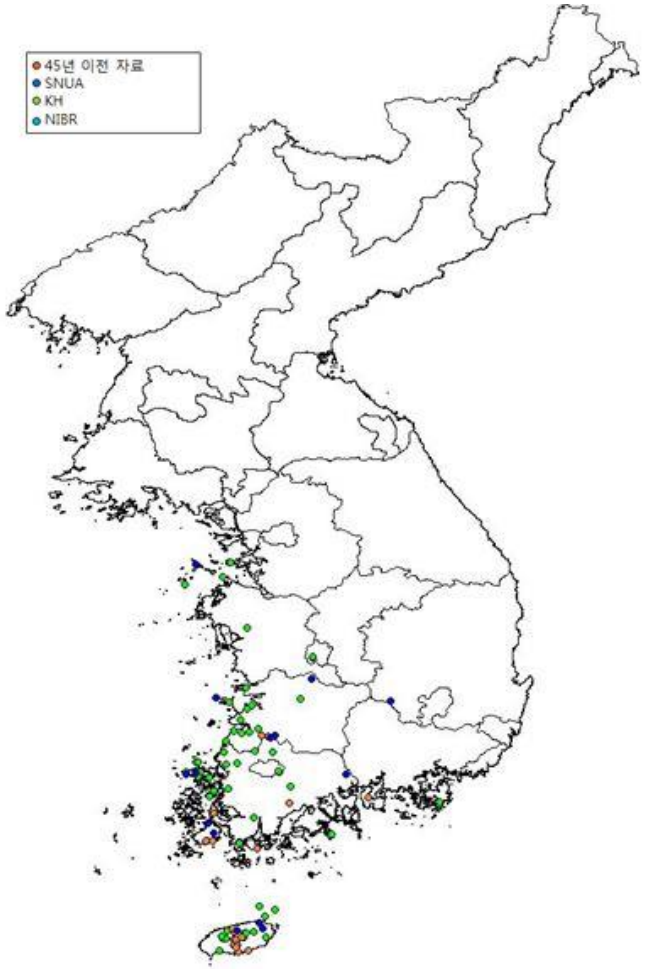
Group2. *Osmanthus heterophyllus* (G.Don) P.S.Green



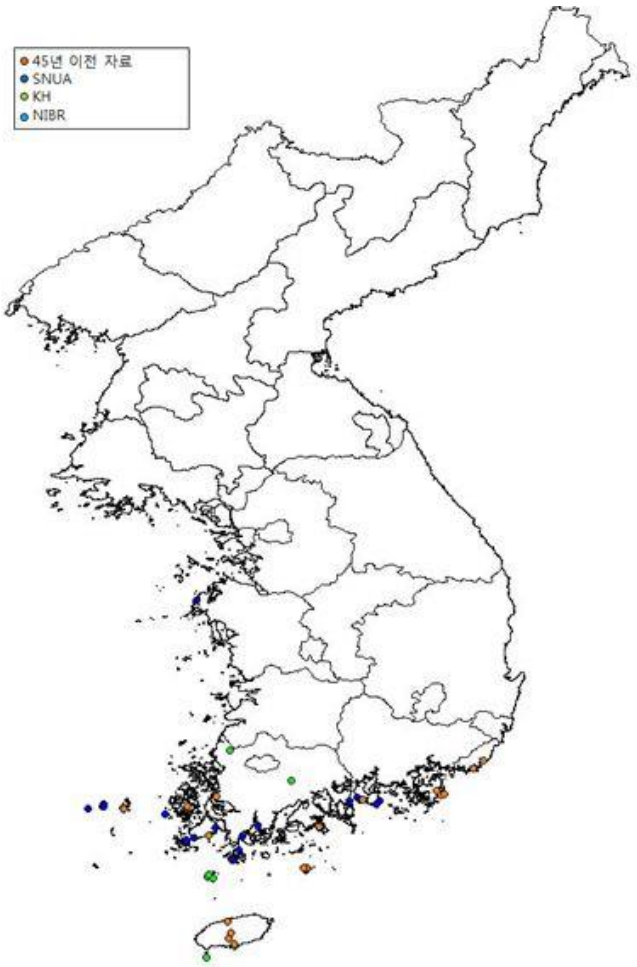
Group2. *Aphananthe aspera* (Thunb.) Planch.



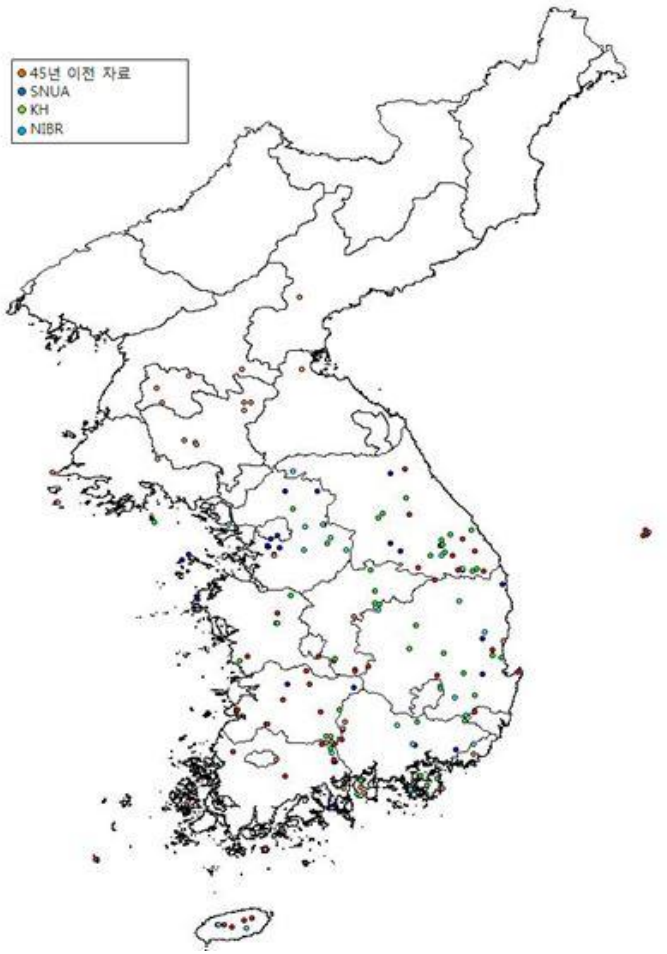
Group2. *Ostrya japonica* Sarg.



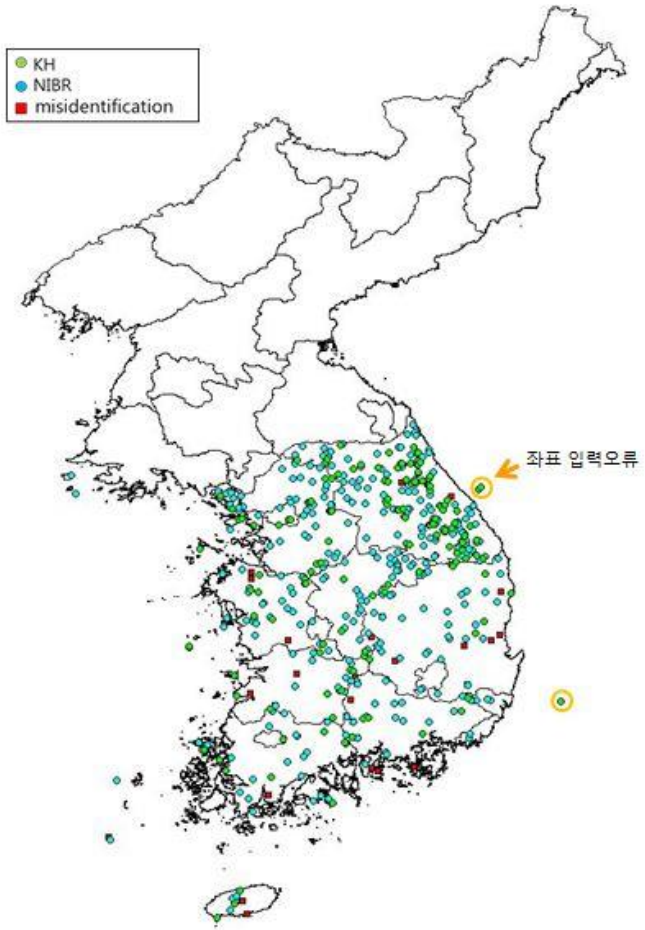
Group2. *Orixa japonica* Thunb.



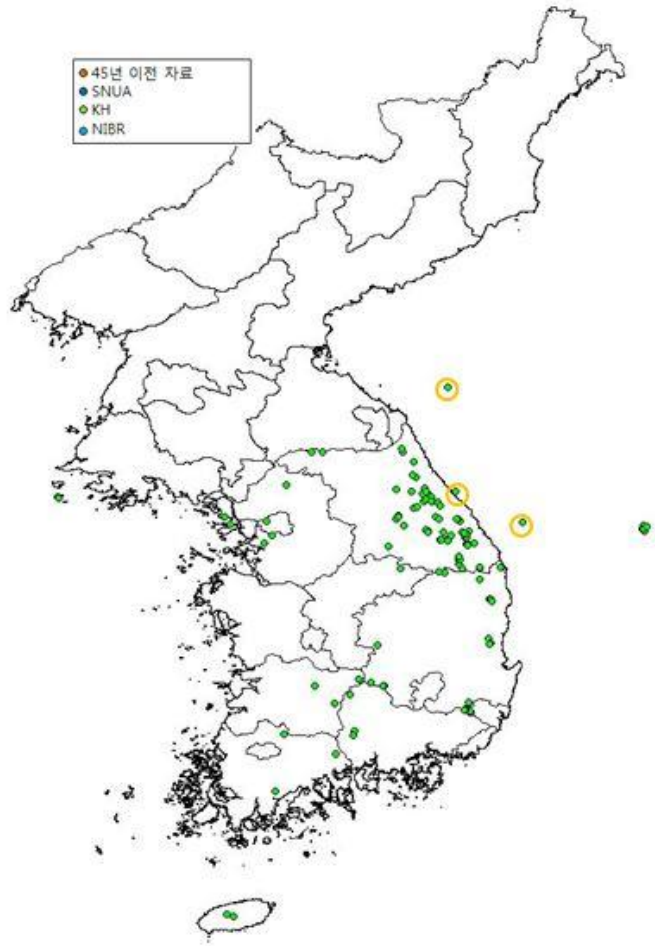
Group 2. *Rhaphiolepis indica* var. *umbellate* (Thunb.) H.Ohashi



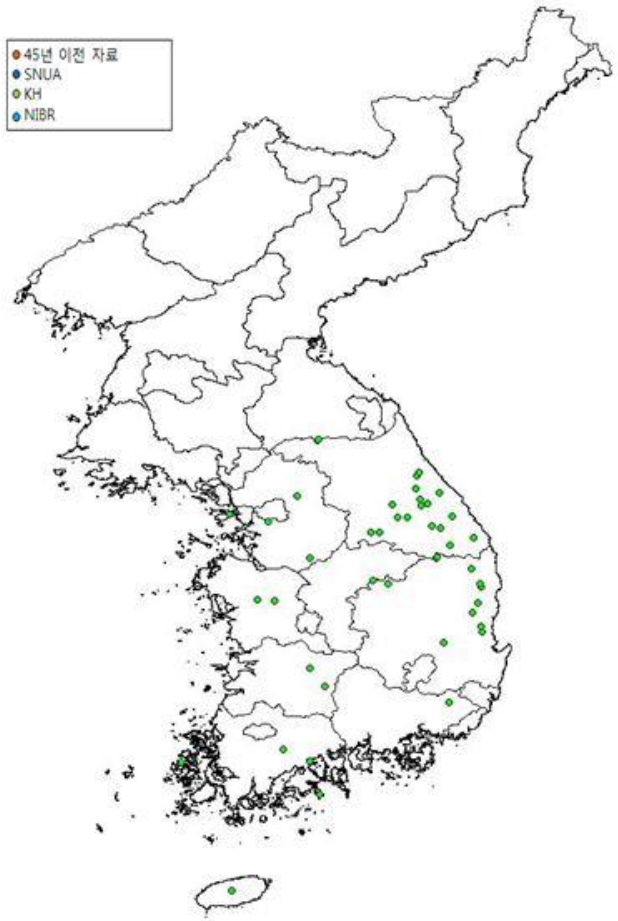
Group3. *Cornus walteri* Wangerin



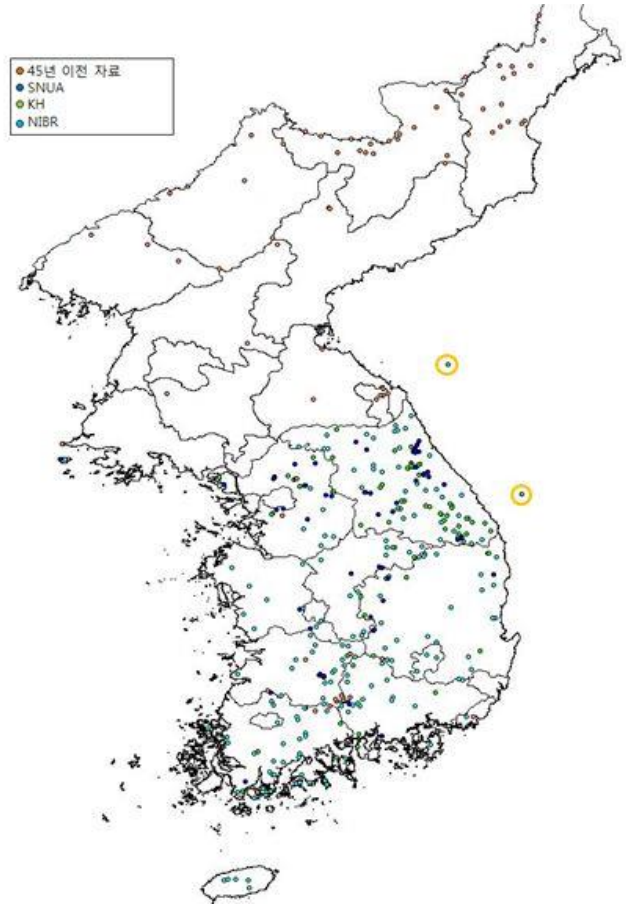
Group3. *Fraxinus chinensis* Roxb. var. *rhynchophylla* (Hance) Hemsl.



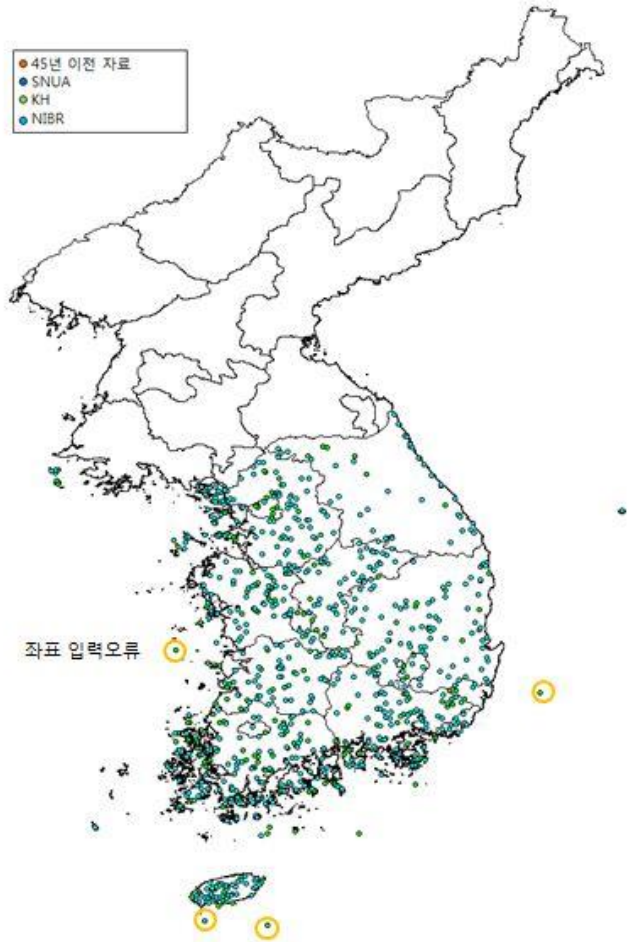
Group3. *Tilia amurensis* Rupr.



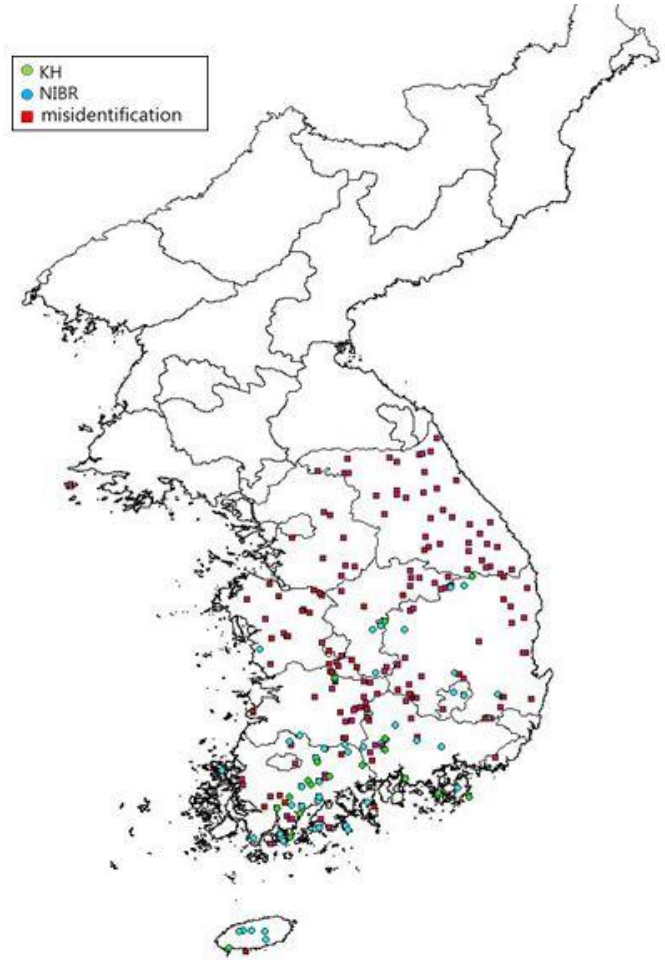
Group3. *Tilia mandshurica* Rupr.Maxim.



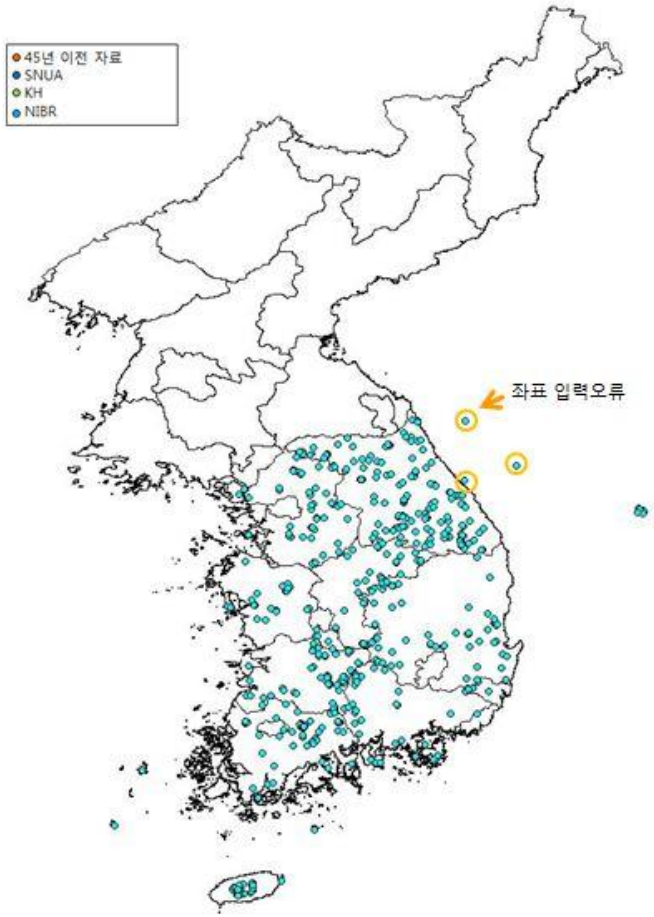
Group3. *Corylus sieboldiana* Blume var. *mandshurica* (Maxim.) C.K.Schneid.



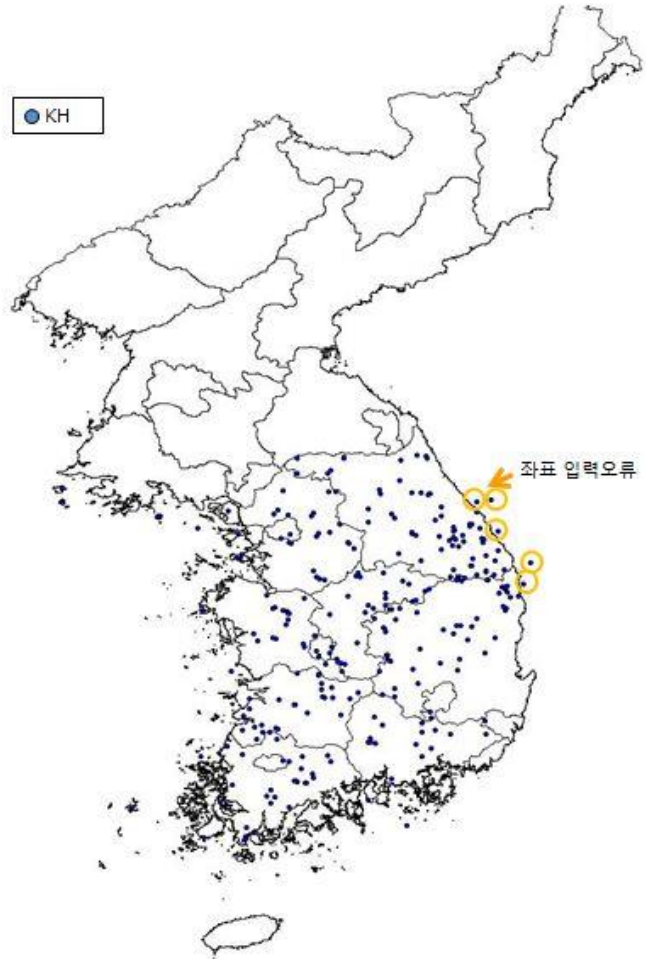
Group3. *Ligistrum obtusifolium* Siebold & Zucc.



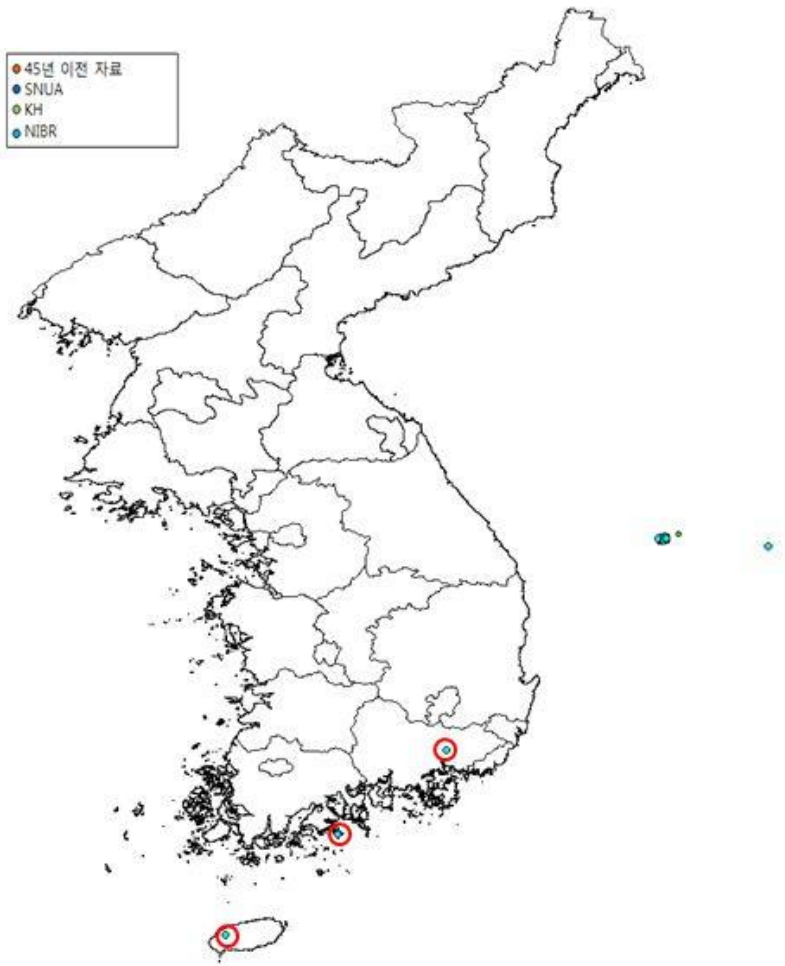
Group3. *Corylus sieboldiana* Blume



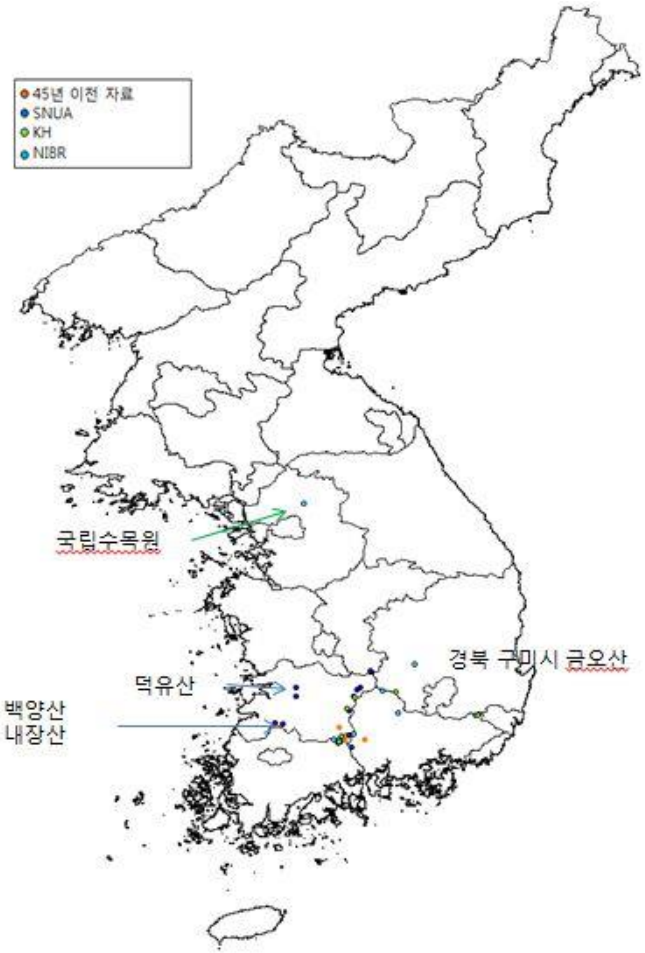
Group3. *Cornus controversa* Hemsl.



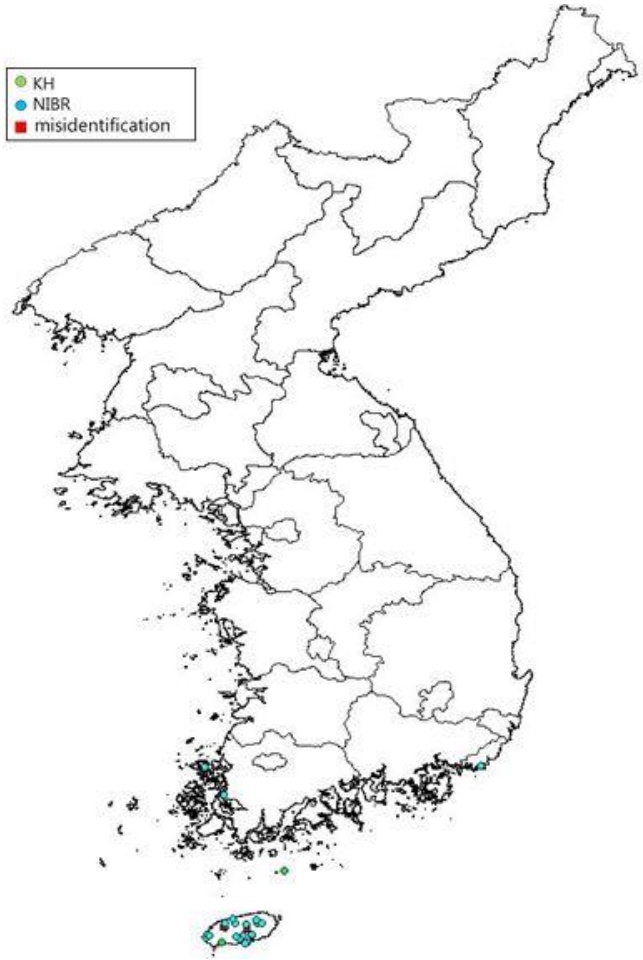
Group3. *Flueggea suffruticosa* (Pallas) Baill.



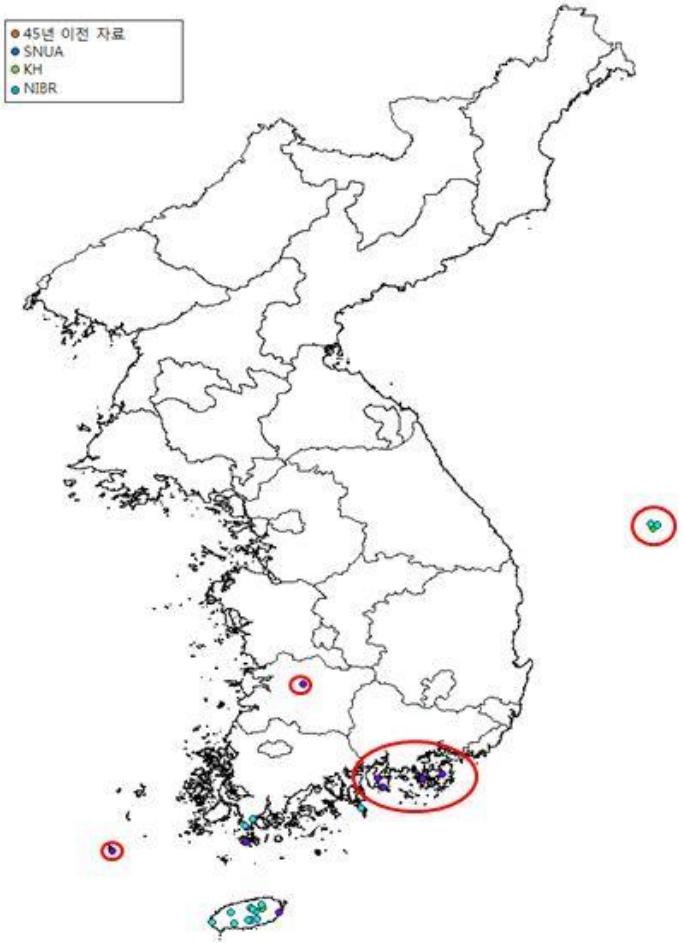
Group4. *Ligustrum folisum* Nakai



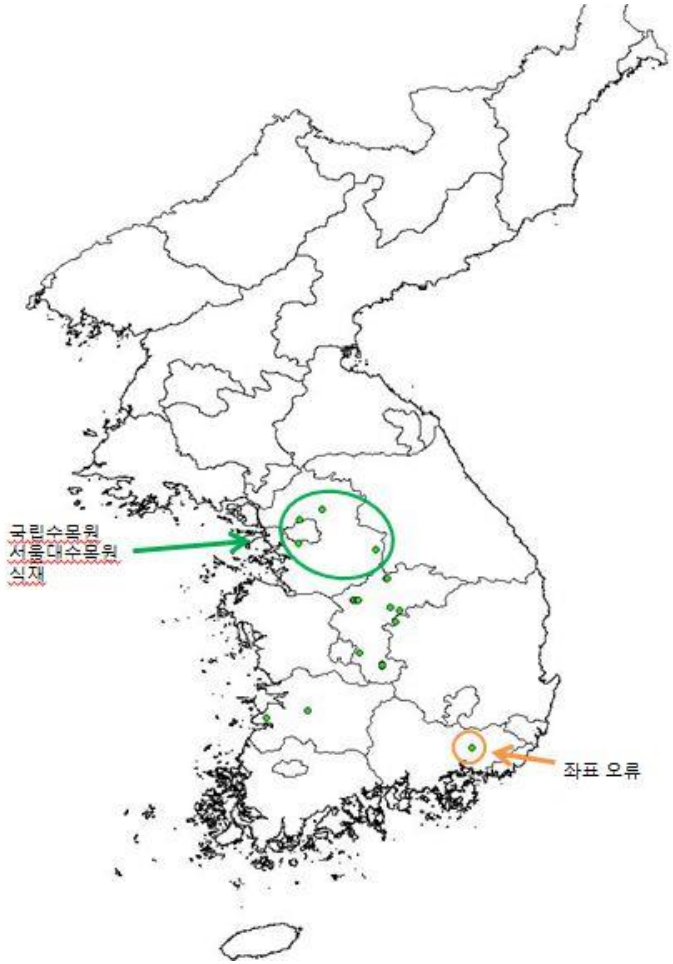
Group4. *Fraxinus chiisanensis* Nakai



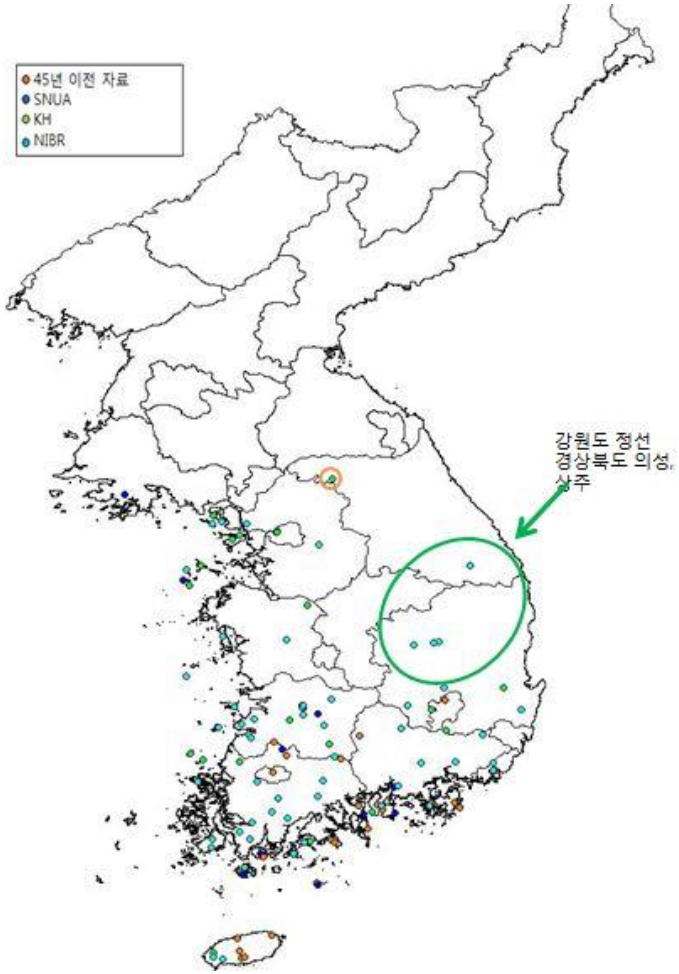
Group4. *Ligustrum lucidum* W.T.Aiton



Group4. *Ligustrum ovalifolium* Hassk.



Group4. *Abelliophyllum distichum* Nakai



Group5. *Chionanthus retusus* Lindl. & Paxton

Abstract

Herbarium holds specimens that represent the information of hundreds years and has been provided valuable information for biodiversity researches. With advances in information technology, primary occurrence information has been digitized and become more important as the source of enormous bioinformatics data. In Korea, National herbaria have created digitized data in the past 15 years. For this effort to work, limitations such as lack of qualified taxonomic determination, precise georeferencing of the data and updated taxonomic treatment should be overcome. This research tried to (1) assess the current status of plant specimens in National herbaria (2) infer the cause of errors by analyzing the type of misidentification (3) suggest the process for detecting and cleaning misidentified specimens. Data were kindly provided by the Korea National Arboretum(KH) and the National Institute of Biological Resources(KB) and transferred to BRAHMS database. 17,517 herbarium records of woody plant families, mainly Oleaceae and other 10 families were used for analysis. All the specimens were examined and determined by visiting both herbaria. The rates of misidentification were various from 0% to 67.07% by taxon and there was rare association between the rate of misidentification and

taxon which could decrease the reliability of the data and cause misunderstanding and misuse of the data. This research tried to detect errors using database of specimen before 1945 and woody plants of Korea site(http://florakorea. myspecies.info/en). By using this reliable distributional information and comparing geocode, taxonomic errors and spatial errors can be cleaned in advance. In the aspect of DB management, 30-38% of disagreement was found when comparing KH DB with specimens in the Herbarium(*Fraxinus* and *Syringa*). The key of database management is that all the process regarding specimens should be executed through the database.

Keywords : Herbarium specimen, Herbarium records, Primary occurrence information, Data quality, Data cleaning, taxonomic misidentification, detecting errors

Student number : 2015-23017