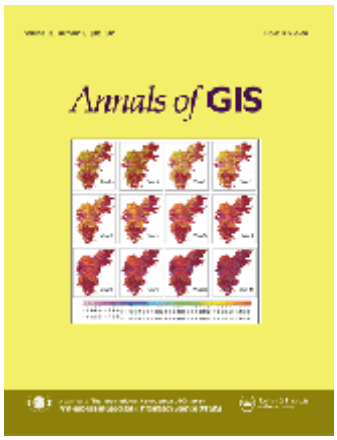

Disentangling the influence of geographical laws and sampling biais to model distribution of Birch from Openacess Dataset in Swedish Lapland

Journal:	*Annals of GIS*
Manuscript ID	TAGI-2020-0067
Manuscript Type:	Research paper
Keywords:	Biogeography, Geographical Laws in Spatial Ecology, Biodiversity Databases and Datasets, GBIF, Birch tree, Swedish Lapland

Disentangling the influence of geographical laws and sampling biais to model distribution of Birch tree from Open-acess Biodiversity Dataset (OBDs) in Swedish Lapland

Abstract—**The current biodiversity crisis, combined with climate change are major issues requesting specific monitoring of plants communities' responses in terms of geographical distribution. Nowadays, large open-access biodiversity datasets (OBDs) such as the Global Biodiversity Information Facility (GBIF) are commonly used to describe, explain and predict fauna and flora geographical distribution. They constitute new opportunities, but stay related to major uncertainties about sampling biases, driven by the concentration of various biodiversity data records and associated data providers. Taking the example of a widely studied tree (*Betula pubsescens* Ehrh.), in a scientifically well-funded region (Swedish Lapland), we discuss those peculiar issues in the frame of geographical laws (e.g. respectively spatial autocorrelation, heterogeneity and similarity) at macro- and micro- regional scales. After spatial and temporal filtering on georeferenced records and discussion on sampling strategies heterogeneity, tests of spatial autocorrelation (Moran's I index) has been conducted on Birch tree records provided by major institutions, comparatively. Pearson Khi-2 (χ^2) test is thus applied on the generated grid to confront number of Birch tree records with accessibility factors (e.g. artificial land cover, roads, protected natural areas) Thus, a micro-regional analysis is conducted to quantify Birch tree records in vegetation classes where this tree species is supposed to be dominant. At the macro-regional scale, results show the high spatial variability of sub-datasets according to institution providing records from the studied GBIF OBDs (with higher autocorrelation results for large contributors). This spatial variability, and high spatial autocorrelation effects appears to be partly explained at macro-, micro- and local scale by the distribution of human accessibility and facility factors (e.g. roads, cities etc). In this study case, exploring OBDs for an extensively addressed tree species, in a significantly funded region/country was particularly useful, demonstrating the relevance of spatial similarity law to differentiate adequately sampling biases efforts from "natural" spatial autocorrelation.**

Keywords— Biogeography; Geographical Laws in Spatial Ecology; Biodiversity Databases and Datasets; GBIF; Birch tree; Swedish Lapland; Habitat Niche Modelling, Climate Change

I. Introduction

A. *Geographical distribution of species, geographical distribution of sampling biases : deciphering tracability of Open-access Biodiversity Datasets records with general principles in geography*

At a time when there is an urgent need to model and simulate geographic space, and at the same time, a large amount of data is available online to feed these models, the geographer has to ask himself the following question: how to disentangle the influence of geographic laws and sampling bias related to these large open-source data and avoid misinterpretation?

This is particularly true for the modelling of vegetal species distribution, particularly for food species bringing important ecosystem services for the society. The distribution range of these species is expected to move with climate change, with dramatic changes for socio-ecosystems, interacting with biodiversity crisis (Pearson et al., 2013), particularly in arctic and subarctic regions (Pfeifer et al. 2019, IPCC 2013, IBPES 2019). The current global environmental crisis constitutes an urgent need to model and simulate geographic space to get prepared and seeking for resiliency at all geographical levels. Meanwhile, publicly available geographical data about environmental conditions and biodiversity (also known as Primary Biodiversity Data, hereafter named Open-access Biodiversity Datasets, OBDs), Anderson, Araùjo et al., 2020) are currently easily and freely accessible, coming from various data aggregators such as GBIF (2016), DataDryad, IUCN, Neotoma, NatureServe, Conservation of Arctic Flora and Fauna, CHELSA-Project, WorldClim, etc (Sillero & Barbosa, 2020).

Under these conditions, it is necessary for geographers and other disciplines interacting with space management (biology, ecology, engineering, economics; political and law sciences etc) to question and to put in perspective the reliability of such easy-to-process datasets and associated software (Sillero & Barbosa, 2020; Zizka, Antonelli et al., 2020). Without mentioning the influence of spatial correlation and covariation between species locations, and interpolated abiotic factors/datasets used to model patterns of distribution or migration, it is thus particularly crucial to sum on to the influence of geographic laws in sampling bias related to OBDs and having hindsight on those datasets, quantitatively important in georeferenced records (~10^9 individual species records, Andersen, Araùjo et al., 2020) but qualitatively peculiarly questionable when modeling those individuals, taken without any ponderation when processed (Sillero, Barbosa, 2020). The other side is also true, three mixed groups contributing to nourish, aggregate and use OBDs (data providers, data aggregators, data users), with little respect (or no time) to carefully providing guidance on first-hand biodiversity datasets, which will be later used for a numerous of usages (Andersen, Araùjo et al., 2020).

In an absolute way, OBDs mainly consists on georeferenced records of specie, which are in a first sight one location of one individual belonging to one taxa (fauna, flora, fungus, etc). Lack of information regarding metadata (e.g. year, data acquisition conditions, sampling efforts, and research questionings) constitutes gaps and biases, which are common for OBDs. They can be summarized in five interlaced types (El-Gabbas & Dormann, 2018; Zizka & Antonelli, 2020):

- Biological biases (taxonomic bias represented by biologist/ecologist selectivity about species of interests, taxonomic misidentification, quality monitoring not being provided)

- Spatial biases (remote areas and proximity effects; errors when geo-referencing/inaccuracies and approximations on field and hereafter; georeferenced collections reflecting collection/museum location instead of where those occurrences have been found; "road-side-bias" effect and more widely physical accessibility and)
- Environmental biases (harsh environments, "avoided" studied areas because of their difficult conditions for researchers)
- Temporal biases (absence of temporal information; inaccuracy, seasonal biases when sampling)
- Socio-economic and political biases (sampling biases linked to research history and funds, conflicts, human rights)

A striking example of multi-factorial sampling bias at a global scale is the more important number of GBIF records in well-financed countries, compared to species-rich countries (El-Gabbas & Dormann, 2018).

Models and their cross-validation, and more deeply first-hand datasets as input of models, such as available in OPDs constitutes important opportunities which should be used in the framework of spatial distribution laws and documentation on their use in such framework disseminated (Cotello et al., 2013 in Anderson, Araujo et al. 2020). Another important gaps of knowledge found in the literature is represented by the "*few attempts* [which] *has been made to compare the geographic sampling bias among datasets* […] [to quantify] *the effect size of specific bias factors and compare it among them*" (Zizka, Antonelli et al., 2020).

B. Birch tree (Betula pubescens Ehrh. 1791), a key-stone species of subalpine woody tundras

Biogeographically, the Birch tree (*Betula pubescens* Ehrh. 1791) is widespread across Eurasia in mountainous areas of the boreal zone (Tutin, Burges et al., 1993). According to Andersson (2005), *Betula pubescens* is present in Swedish Lapland with two subspecies: *Betula pubescens* ssp. *pubescens*, named Birch tree and *Betula pubescens* (Ehrh) ssp. *tortuosa (*Ledeb), named mountain Birch. Previously they were considered as two different species named *Betula pubescens* Ehrh. and *Betula czerepanovii* n.i. Orlova). The first one is significantly present in the boreal forest, in monospecific stands or as subdominant species in conifer forests and more generally in Middle and Northern Eurasia. The second one is present on the North-Western Atlantic fringe of the distribution map of Betula ssp., at the foothills of the Fennoscandian chain. Unfortunately, in the OBDs used in this paper, the information about subspecies is available for only 8% of the records. Subspecies cannot be distinguished at the regional scale of Swedish Lapland, but only Birch tree is represented at the micro-regional and local scales of our study (section II.A).

The high genetic variability of mountain Birch renders it particularly flexible, although the species is hygrophilous and prefers clay, explaining its affinity with peats, moorlands to be later colonized by it. Patchworks of wood Birch tree woods are a-zonal subalpine tundra ranging from low to middle-elevation in northern Europe and Siberia. Indeed, the tree species and associated floristic communities form an ecotone between northern boreal forests and scattered, perhaps vascular-free tundra at high altitude and/or high latitude (Rydin, Snoeijs et al. (eds), 1999). Along the northern boreal zone in Eurasia, *Betula* genus are often considered as the last trees to survive to such cold conditions, and form then tree-lines. Nowadays, observational and experimental studies in palynology, genetic and ecology have demonstrated the high variability of the Birch tree-line, and its high potential to invade higher altitudes, in partly due to the direct effect of warming on recruitment (Kullman, 1993; Truong, Palmé et al., 2007; Bryn & Potthoff, 2017; Kullman, 2017).

This present study, focused on *Betula pubescens*, including its two subspecies, aims to better assess the present and future distribution of the species, being the interlaced links between tree biomass and palatable vegetable for reindeer and for reindeer herding systems. Indeed, even if young leaves of mountain Birch might be consumed by subarctic and arctic herbivorous, and particularly semi-domesticated reindeer (Forbes & Kumpula, 2009), impacts of invasion and densification of Birch tree woods on biotopes and other biological communities might be important (Courault, 2018b; Agnan, Courault et al., 2019). Although direct inter-specific interactions are complex between Mountain Birch and reindeer, tree-line rising, biomass densification and phenological aspects of the tree are hypothetically susceptible to interfere on reindeer biological rhythm and migration in spring and summer (Kumpula, Stark et al., 2010). Beyond that, it is important to survey and assess the species distribution in the past, present and future, among which habitat modeling niche techniques could be useful to infer and predict on its future distribution at different ecological and geographical scales.

C. Geographical laws in the frame of Openaccess spatial ecology : questioning and objectives of the study

Spatial autocorrelation is a property of the First Law of Geography, Tobler's Law 'Everything is related to everything else, but near things are more related than distant things' (Tobler, 1970 in Sillero & Barbosa, 2020). Auto-correlation is necessary for modeling descriptive and explanatory species distribution areas (Segurado et al., 2006, Dormann 2007, Dormann et al. 2007, De Marco et al., 2008 in Sillero & Barbosa, 2020), but misleading because of biases and quality heterogeneity of OBDs. Nevertheless, for ecologists and biogeographers, spatial autocorrelation might appear to a vague concept, often confused with spatial clustering of records and partly diluted with filtering records (Sillero & Barbosa, 2020). Hence, spatial clustering might appear to be either a result of 1) data aggregation of various sources 2) sampling efforts coming from different data providers 3) species' distribution in its natural habitat.

The second point linking biodiversity research, such as habitat niche modeling, and geographical laws is the essential bridge between biotic and abiotic relationships in space. Strong environmental gradients are theoretically related to more realistic models (e.g. increased explanatory or predictive power of the model (Seoane et al., 2005 in Sillero & Barbosa, 2020), based on the geographic similarity principle, corresponding to the "contagious model" opposed to the "random model" (spatial heterogeneity) of species distribution by Gounot (1969). Indeed, Varela et al. (2014) showed that filtering by environmental criteria provides better results (Sillero & Barbosa, 2020). This justify the study area as well as the studied species, *Betula pubescens*.

Into this conceptual framework, the case study is based on Birch tree (*Betula pubescens* Ehrh.), whose biological/ecological characteristics are relatively well studied and geographically enquired at the scale of its current distribution. Thus, our hypothesis is that spatial biases coming from OPDs are still detectable and won't provide sufficient and reliable geographical information to firmly describe and predict its present and future distribution using habitat niche modeling techniques. In particular, accessibility and facility factor would be likely to be quantified at various geographical scale, such as macro-, micro- and local levels in Swedish Lapland.

The objective of this paper is to test to what extent and under what conditions OPDs' sampling biases can be confused with geographical laws, and to propose methods to separate their respective influence. We will present the case study, the material and the methods mobilized (section II), then the results obtained at 3 scales, macro-regional, micro-regional and local (section III) and will conclude with a discussion and the methodological perspectives opened for geographers by this study.

II. Study Area, Material and Methods

A. *Macro- and micro- regional study areas*

The macro-regional study area, represented by the Swedish Lapland (e.g. *Sameby* in northern Sàmi) is complex ecologically, historically and geographically (Forbes & Kumpula, 2009; Ojala, 2009, Löf, 2013). Administratively speaking, the macro-regional studied area stretches from Svealand to Norrland old Swedish provinces, and comprise as well some Norwegian territories (= 51 reindeer herding communities, e.g. *samebyar*). Indeed, the studied area corresponds to the Swedish part of the Lapland whose ecosystems and landscapes were economically, culturally and socially constructed and valued by Sàmi people, and fiercely negotiated with Scandinavian authorities along the last centuries (Manker, 1953; Svonni, 2010). Having Finno-Ugrian origins, linguistic heritage and culture, cultural landscapes have been (and are still) mainly devoted to migratory reindeer herding, with a specialization of pastoral system according to clans and families' belongings. Landscapes of Swedish Lapland, as synoptically represented in the Figure 1, present a diversity according to their location within 4 main gradients/ecoclines: latitudinal, longitudinal, continental, altitudinal (Courault, Duval et al., 2018, Courault, 2018). More precisely, vegetation zonation of Swedish Lapland ecosystems expands from the Southern Boreal zone, to Middle- and Northern Boreal subzones, to Subalpine and Alpine belts compose landscapes of the Fennoscandian mountains, devoted to calving areas and summer pastures for reindeers. Zonal as well as a-zonal vegetal landscapes are highly dependent on the climate (colder above the Arctic Circle, winder and milder in glacial valleys influenced by the Gulf Stream), mineral soils (and associated bedrocks), hydrology and (micro-) topography (Rydin, Snoeijs et al. (eds), 1999). At this level, describing the current and future distribution of birch tree (*Betula pubescens* Ehrh.) is particularly interesting in both fundamental and applied research in biogeography, being the last "tree" to remain at the Subalpine belt and the frame of current climate change.

The micro-regional study area, as displayed in the Figure 1, is located in the northern edge of the Swedish Lapland. Four *samebyar* (reindeer herders' communities) are represented within this area, and are characterized by mountainous reindeer herding pastoral systems (e.g. wider vital areas of semi-domesticated reindeers, wider activities of owners, and wider contemporary disturbances; Manker 1953; Courault, Duval et al., 2018). In terms of vegetal landscapes, this area is particularly interesting because of its transitional situation between the Boreal zone, where Scots pines are highly exploited, as well as the alpine and subalpine belts, where winter and summer tourism might constitute troubles and tensions for reindeer pastoral systems as well as for local biotic communities (Skarin, Danell et al., 2008). This micro-regional study area has been selected for multiple reasons, because of the naturalist entanglement between "traditional" activities, modern ones, juridical issues in terms of environmental protection and wildlife watching by scientists (INTERACT-network of natural science stations). Furthermore, the area has been chosen because where field work carried out by Courault (2018) provides particular expertise on OBD quality and Betula genus is represented by Mountain Birch (*Betula pubescens* ssp *tortuosa*)

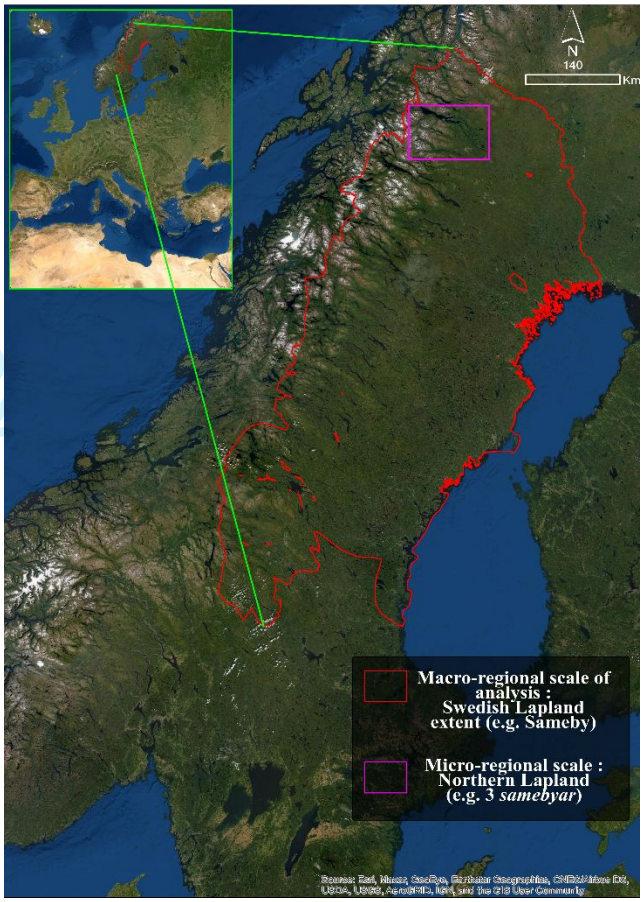

Fig. 1. Sudied areas in Sweden: macro-regional scale (Swedish Lapland, *Sameby*), and micro-regional scale of analysis of mountain Birch records from GBIF dataset (Sources: Swedish Saami Parliement, RenGIS 2.0 sametinget.se; Lantmäteriet, SLU; realization R.Courault 2020)

B. *Material*

1) Acquisition of georeferenced Birch tree records, semi-automatized filtering and study area scaling

Using the *Rgbif* package (R Studio), georeferenced occurrences of Birch tree (Betula pubescens Ehrh.) were downloaded at the scale of Eurasia (e.g. 100 000 records; GBIF). For spatial biases which are analyzed here, cleaning records (e.g. removing removing and keeping occurrences of

one or several species) using distances and uncertainties of their locations appears to be one solution "*to dilute the effect of uneven sampling effort across the study area*" (Anderson and Raza, 2010, Boria et al., 2014 in El-Gabbas & Doorman, 2018). With the *Clean Coordinate* package (R Studio) spatial tests were applied to flag and to remove:

- Record without any coordinates
- Records belonging to museum' and institutional' collections, herbarium and revealing their location instead of the exact place where records were sampled (non-human in-situ observations)
- Non accurately georeferenced records, be it for national capital, province/administrative region centroids
- Geographical locations with more than one record
- Spatial uncertainty of coordinates being above the threshold of 1000 meters (Figure 2).

Once spatial tests conducted using R packages, georeferenced occurrences of Birch tree has been imported in Qgis 3.14.16 to intersect Eurasian occurrences at the extent of Swedish Lapland, considered as the extent of all Swedish reindeer herding communities' borders (source: sametinget.se; RenGIS 2.0).

Thus, after dataset visual control (Figure 2), an attribute query deleted georeferenced Birch tree being recorded before 1995, until today (2020). This temporal filtering has mainly ecological/biogeographical goals in terms of habitat niche modeling. The preprocessed dataset gives then 12 265 records (Table 1).

TABLE I. NUMBER OF RECORDS BEFORE SEMI-AUTOMATIZED SPATIAL, BIOLOGICAL AND TEMPORAL FILTERING OF BIRCH TREES, GIVEN BY INSTITUTION PROVIDER AT THE MACRO-REGIONAL SCALE (SWEDISH LAPLAND)

Institution source/provider	**Σ records before semi-automatized spatial and biological filtering (1995-2020)**	
	Σ records	***%***
Artdatabanken	9831	80.1
Calluna AB	50	0.4
Ecofact	3	0.02
Gävleborgs Botaniska Sällskap	986	8
Jämtlands Botanika Sällskap	25	0.2
Länsstyrelsen Jämtland	404	3.3
Länsstyrelsen Norrbotten	15	0.12
Länsstyrelsen Västerbotten	102	0.83
Länsstyrelsen Västernorrland	81	0.7
MFU	2	0.02
NA	2	0.02
NTNU-VM	7	0.06
NV	743	6.1
O	11	0.09
Piteå kommun	3	0.02
TOTAL	12265	100

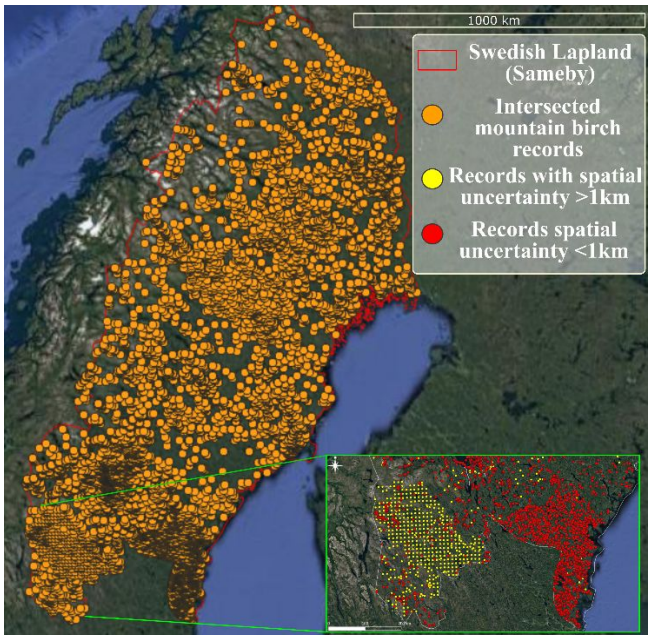

Fig. 2. Example of data visualization and balance between sampling effort (on field) of Birch tree GBIF dataset, its geographical uncertainties on the one hand and statistical needs for records and niche species modelling on the other hand. Note the systematic and squared sampling effort/strategy, in the southwestern part of the study area, presenting a significant spatial uncertainty (e.g. 2.5km , institution: ArtDataBanken). Nevertheless, this systematic sampling minimize the bias due to presence only data set, the species absence is noticeable in two areas. Even if overall autocorrelation (scale of the study area) might be over-represented, geographical information missed by an arbitrary cleaning constitutes a serious compromise to take into account in ecological modeling. udied areas in Sweden: macro-regional scale (Swedish Lapland, *Sameby*), and micro-regional scale of analysis of Birch tree records from GBIF dataset (Sources: GBIF, Swedish Saami Parliement, RenGIS 2.0 sametinget.se; Lantmäteriet, SLU; realization R.Courault 2020)

2) Acquisition of geographical data on land cover, accessibility and facility factors

The portal Copernicus – Land Monitoring Service (https://land.copernicus.eu/pan-european/corine-land-cover) has been used to acquire the Corine Land Cover map of 2018 (CLC 2018), in its shapefile format. When pasting the CLC 18 at the macro-regional extent of the study area, an attribute query is made to select land cover categories related to artificial zones and human facilities, excluding other land covers from the analysis:

- (1) Continuous urban fabric
- (2) Discontinuous urban fabric
- (3) Industrial or commercial units
- (4) Road and rails networks and associated lands
- (5) Port areas
- (6) Airports
- (10) Green urban areas
- (11) Sport and leisure facilities

Thanks to the SLU and Lantmäteriet, *Vegetation fjällkedjan vektor* (shapefile of mountainous vegetal communities) and *Vegetation Norrbotten vektor* (shapefile of boreal vegetal communities of the Norrbotten province) were downloaded (https://www.lantmateriet.se/) and merged as geoprocessing/

Concerning Natural protected areas, Nationally Designated Areas (CDDA) have been downloaded from the European Environmental Agency Open Access data portal (https://dd.eionet.europa.eu/datasets/latest/CDDA) to get conservation areas of Swedish Lapland.

For roads and railways, divagis.org has been used to get Norwegian and Swedish transportation networks in shapefile (https://www.diva-gis.org/gdata).

The last two datasets (Natural protected areas and roads and railways) have been intersected with the extent of the Swedish Lapland, and merged in particular for transportation networks.

Figure 3 shows that human observations of Betula pubescens are concentrated in the boreal forest, but extend to the South-western Botnia Gulf (Birch tree) and to the North-western foothills and valleys of the Fenno-Scandian chain (Mountain Birch). Geographical bias effects are visible at the scale of a territory. For those recorded human observations of Birch tree, concentration effects are noticeable, e.g. Norrbotten county and around Östersund, respectively in the northern and southern parts of the macro-regional study area.

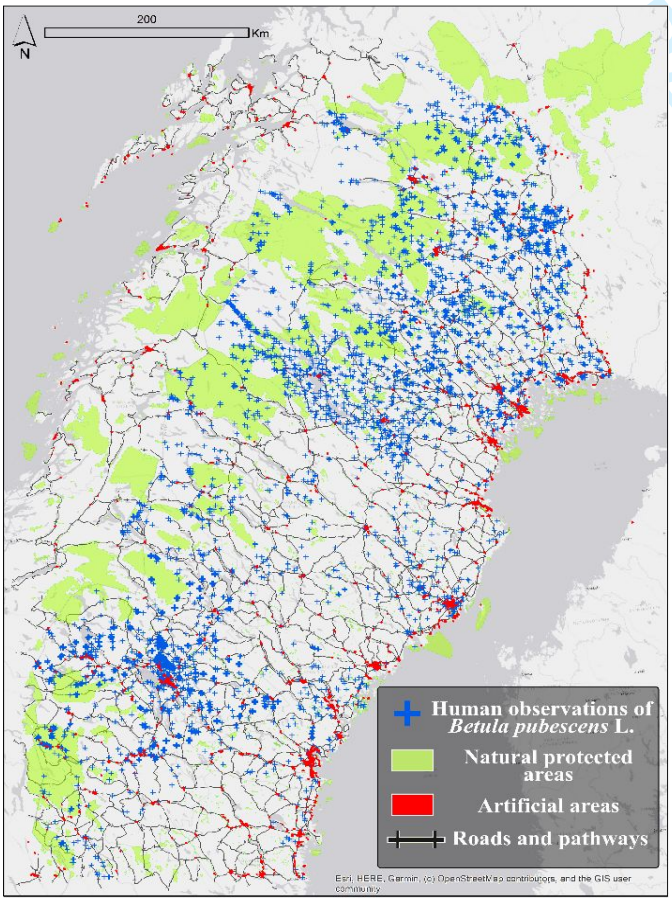

Fig. 3. Cartographic representation of filtered records of Birch tree and tested accessibility/facility factors (Sources: GBIF, Corine Land Cover 2018 Copernicus, European Environmental Agency and DivaGIS; realization R.Courault 2020)

C. Methods

1) Spatial and statistical analysis at the macro-regional scale

For the macro-regional scale and analysis conducted on Birch tree records and tested geographical features (autocorrelation, accessibility and facility factors), grids computing has been chosen for several reasons. Firstly, grid cells avoid statistical noise which could be due to pre- or post-processing, such as interpolation methods (Ruete, 2015). Secondly, it technically simplifies counting number of occurrences of both studied phenomenon (here Birch tree) and explanatory factors (its distribution among institution providers, as well as accessibility/facility factors which could potentially explain Birch tree distribution as recorded by GBIF).

a) Grids generations and absence/presence countings

Using the tool "*Create grid*" from Gdal on QGis 3.14.16, grids were generated at the spatial extent of each dataset corresponding to institutions records of the Birch tree, at the wider scale of the study area (e.g. Swedish Lapland). As represented in figure 4, the wider grid used to count records of Birch tree and associated factors is displayed by the Artdatabanken providing institution. Other generated grids correspond to the extent of each providing institution. To include the modal spatial uncertainty given by records in the GBIF dataset, spatial resolution of grids has been set up to 1*1 km for every grid (see Table II for number of columns and rows; projection system of the GIS: SWEREF 99 TM).

For such macro-regional level analysis (autocorrelation test and influence of accessibility/facility factors) those counting include as well other geographical factors as listed in the Material section: artificial land cover (CLC 18), Natural protected areas, roads and railways.

b) Moran's I spatial autocorrelation test per providing institution

Morans'I spatial autocorrelation tests have been conducted by differentiating records of Birch tree *Betula pubscens* Ehrh. from various institutions sources. As shown in the table I, number of occurrences greatly vary accordingly. In this way, only institutions having recorded more than 20 occurrences of Birch tree are compared in Moran's I spatial autocorrelation test. Those are listed below and displayed in figure 4:

- Artdatabanken (Swedish Agricultural Sciences University – SLU; Swedish Species Information Centre; https://www.artdatabanken.se/en/about-us/)
- Länsstyrelsen Västerbotten (Administrative Board of Västerbotten County)
- Länsstyrelsen Västernorrland (Administrative Board of Västernorrland County)
- Länsstyrelsen Jämtland (Administrative Board of Jämtland County)
- Jämtlands Botaniska sällskap (Jämtland's branch of the Swedish Botanical Society)
- Gävleborgs Botaniska sällskap (Gävle's branch of the Swedish Botanical Society; http://gavleborgsbotaniskasallskap.se/om-oss/)

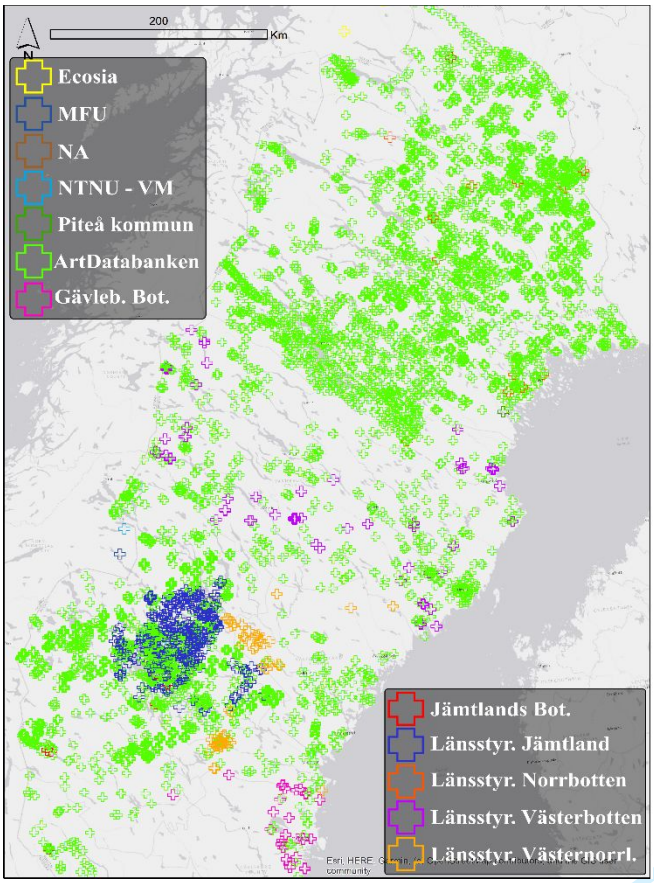

Fig. 4. Distribution of Birch tree records according to the most contributive institutions of the GBIF dataset (Sources: GBIF; realization R.Courault 2020)

Others data providers (e.g. institutions) presenting a low number of records are mainly represented by consult companies, NGO's and one municipality (e.g. ECOSIA, MFU, NA, NTNU – VM, Piteå kommun; Table I, figure 4)

The tool "*Spatial Autocorrelation (Moran's I)*" of ArcMap 10.4 is used to quantify the degree of spatial autocorrelation / dependence between records of the same dataset (e.g records emanating from the same institution) and to compare results between institutions having recorded Birch tree at the scale of the Swedish Lapland.

Because of the polygonal geometry induced by grids (each cell composing grids having a number of records), spatial relation conceptualization parameter has been set to 'Contiguity Edges Corners'; these to better assess spatial relationships between contiguous cells (sides, or cells).

We chose to standardize processed grids and associated counts, to better allow comparisons between Moran's I spatial autocorrelation tests applied for the different institutions. Row standardization is set because generated grids, even though having the same granularity (e.g. 1000 * 1000 m) differ in spatial extents, depending on records distribution/institutional sub-dataset.

c) Chi-2 (χ2) test between number of Birch tree records and accessibility/facility factors

Always at the macro-regional scale (e.g. the entire Swedish Lapland), we were willing to test the degree of dependence between all records (without any distinction of data provider/institution) and factors reflecting scientific/human accessibility and facilities to fieldwork and naturalist observations (e.g. transportation networks, artificial areas such as housings, natural protected areas as an attractive factor, see subpart Introduction).

We used the grid generated at the level of all Birch tree records (e.g. grid having the most important extent; "*Grids generations subpart*"), which cover the entire Swedish Lapland (e.g. 489 291 cells of 1km*1km in SWEREF-99 TM) to check the number of co-occurrences of Birch records and transportation, protected areas and artificial zones. Once counting done, the attributary table has been exported as a CSV file, and then processed with XLStat 2020 to produce a contingency table on those ordinal dependent / explicative variables. A Pearson Chi-2 (χ2) is then calculated on the contingency table, giving as outputs χ2 by cell, observed and theoretical frequencies, observed, critical and p-values of the overall χ2.

2) Spatial and statistical analysis at the micro-regional and local scales

a) Chi-2 (χ^2) test between number of Birch tree records and vegetal communities categories

At the micro-regional scale (figure 1), we wanted to test the independence hypothesis between records of Birch tree and type of ecological habitat as described by the Swedish ecological map (see subpart Material). Thus, number of occurrences by polygon and vegetation communities' categories have been enumerated and summarized using both Qgis 3.14 and XlStats 2020. A contingency table has been produced for the 44 different habitats. Artificial areas were excluded from the analysis. χ^2 test is then conducted on those categorical variables.

b) Study case of Caluna AB data provider at a local scale with photointerpretation

A study case has been conducted at a local scale in the right bank of the Torneträsk lake, near from the Research Station in Environmental Sciences and the eponymous Natural Park (*Abisko naturvetenskapliga station*, INTERACT network and Swedish Polar Research Secretariat, Kiruna municipality, studied area ~1200 ha). GBIF records of Birch tree come from Calluna AB (Table I), aligned along the railway track, which nearly follows the E10 highway. Contours of the tree formations has been digitized (~970 polygons), apart from one cloudy area standing at the center of the study area (source of the imagery: Google Earth, exploited by ArcGIS Pro 2.6). Two subareas, located respectively to the west and to the east, of comparable surface area (~600 ha) are then analyzed by 1) calculating number of Birch tree and their distance in meters to the railway line; frequencies of records by category of photo-interpreted land covers (e.g. Birch forests, water bodies, snow patches) and associated percentages of records by hectares of forests.

III. Results

A. *Spatial and statistical analysis at the macro-regional scale*

a) Moran's I spatial autocorrelation test per providing institution

Table II summarize Morans' I spatial autocorrelation tests by institutional providers having quantitatively contributed to the GBIF dataset of Birch tree at the studied macro-regional scale. The highest Morans' I index is found for Artabanken records (i=+0,15; z-score = 212; p-value <0.0001). Given the

geographical distribution of Birch tree records (figure 3 and 4), two main concentration effects can be detectable, the first almost corresponding to the borders of the Norrbotten county, the second cluster covering the administrative borders of the Jämtland county. The last one condense a lot of records, as confirmed by the second highest Morans' I index for Länsstyrelsen Jämtland (records provided by this institution being in blue, figure 4). Thus, in the southern part of Swedish Lapland, effects of the accessibility on sampling biases appear to be particularly strong, records being intensively clustered around the administrative capital of the county, Östersund. More broadly, other providing institutions might have clustering effects quantified by Morans' I test, although being too weak or even non-significant (for Morans' I test, z-score, or p-value, see Table II).

TABLE II. MORANS' I SPATIAL AUTOCORRELATION TESTS FOR INSTITUTIONAL PROVIDERS SUB-SAMPLES OF BIRCH TREES AT THE MACRO-REGIONAL SCALE (SWEDISH LAPLAND)

Providing institution	Geo-statistical parameters and results					
	Σ columns	*Σ rows*	*Σ records of mountain Birch*	*Moran's I index*	*z-score*	*p-value*
Artdatabanken	565	866	6504	+0.15	212	<0.0001
Gävleborgs Botaniska sällskap	136	92	32	+0.03	7.1	<0.0001
Jämtlands Botaniska sällskap	111	76	25	-0.0002	-0.04	0.96
Länsstyrelsen Jämtland	138	127	392	+0.09	23.9	<0.0001
Länsstyrelsen Västerbotten	329	290	102	+0.03	35	<0.0001
Länsstyrelsen Västernorrland	170	192	81	+0.05	21.9	<0.0001

b) Chi-2 (χ^2) test between number of Birch tree records and accessibility/facility factors

As already mentioned, biodiversity records as available in OBD such as GBIF, are often the reflect of sampling strategies and efforts, whose might be biased by accessibility and facility factors such as transportation, housing, Natural protected areas (El-Gabbas & Dormann, 2018, Zizka, Antonelli et al. 2020). Figure 5 plots Birch tree records counted at the macro-regional extent, cells of the biggest macro-regional grid summarizing number of records of Birches tree, Artificial areas, Road and Railways, Natural protected areas. Quite obviously, number of cells having one or more presence of roads/railways are related to the number of artificial areas as the scattered plot of the Figure 5 shows (upper-right top and lower-left bottom of the Figure 5). For relationship between number of presence of Birch trees and external factors (second column of the Figure 5), presence of artificial areas, roads and Natural protected areas might be classed as follow, given number of presences: 1) presence/number of roads 2) presence/number of artificial areas 3) presence/number of Natural protected areas. This is confirmed by respective standard deviations.

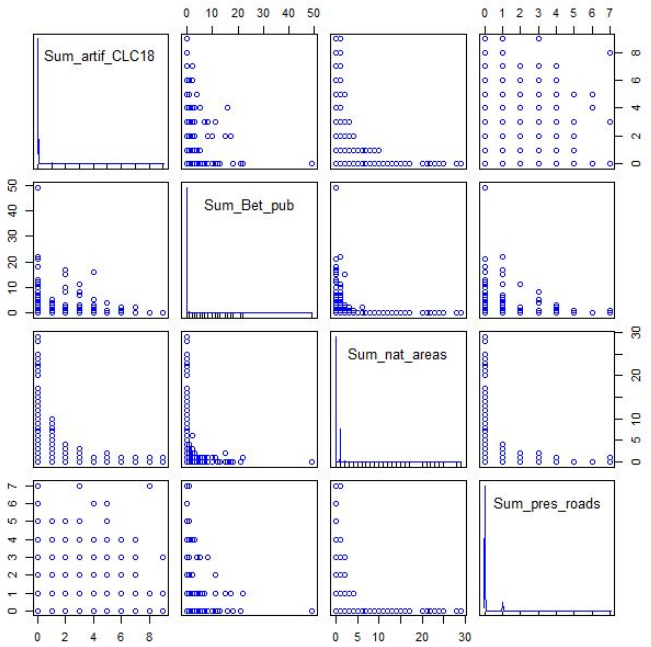

Fig. 5. Plots of Birch trees records counted by cell of 1 km*1 km at the macro-regional scale and number of records of presences of accessibility/facility factors ("Sum_artif_CLC18" : Number of presences of artificial land covers and facilities; "Sum_bet_pub" : Number of presences of mountain Birch; "Sum_nat_areas" : Number of presences of natural protected areas; "Sum_pres_roads" : Number of presences of roads and railways ; realization R.Courault 2020

This is as well confirmed by χ^2 tests applied on this macro-regional dataset, which are highly significant (Table III). Birches trees are significantly over-observed in the grid cells containing a road, an artificial zone or a protected area, in descending order of observed χ^2 value.

The over-representation of birch observations near roads (93% of the observed χ^2 value) can be considered as a sampling bias, as the road facilitates the human observations. However, roads are built in flat valleys on moderately wet terrain to facilitate construction and maintenance and to increase driver safety on straight and level roads. In Swedish Lapland, such glacial valleys correspond to the most frequently common habitat and geographical distribution of the tree.

Birch trees presence in cells containing one or more artificial area (97% of the χ^2 observed value) are over-represented as well (observed $\chi^2 = 485$, critical $\chi^2 = 3.8$, *p-value* <0.0001) and can be interpreted in the same way, both as the result of sampling bias, since observations are more easily made close to an inhabited area, but also because the inhabited areas are located in birch habitat.

On the other hand, the χ^2 value is less important for birch observations in grid cells where there is a protected area (81% of the observed χ^2 value, e.g. 236, critical $\chi^2 = 3.8$, p-value <0.0001). The effect of sampling bias, with an observer preference for protected areas, is not reinforced by a preferential birch habitat effect, unlike the two previous cases.

TABLE III. PEARSON CHI-SQUARED TABLE (X^2) BETWEEN ABSENCE/PRESENCE OF BIRCH TREES RECORDS BY CELL AND PRESENCE/ABSENCE OF ACCESSIBILITY FACTOR AT THE MACRO-REGIONAL SCALE (SWEDISH LAPLAND)

Mountain Birch absence/presence	**Accessibility factor absence/presence**					
	No road	***Road***	***No protected area***	***Protected area***	***No artificialized land cover***	***Artificialized land cover***
No Birch	0.416	6.1	0.381	1.734	0.068	4.3
Presence of Birch	46.02	670.23	42.3	192.2	7.502	473.15
p-value	<0.0001		<0.0001		<0.0001	
Observed χ^2	721.06		236		485	
Critical χ^2	3.8		3.8		3.8	

B. *Spatial and statistical analysis at the micro-regional scale*

a) Chi-2 (χ^2) test between number of mountain Birch records and vegetal communities categories

Table IV displays main results of χ^2 tested conducted between ecological habitat of the Birch trees records and the Swedish vegetation map, at the micro-regional scale of our studied area. For 18 ecological habitats, mainly translated and summarized as peaty soils, covered with conifers, heathland, deciduous forests and/or mosses, Birch trees records are not present. In 21 habitats, Birch tree is observed in less than 4% of the polygons, corresponding to high altitude habitats, be it peaty soils, heathlands and conifer forests.

The three ecological habitats in which Birch trees has been more frequently recorded correspond to "*Moss rich broadleaf forest*" (60 observations in 6% of the polygons), and in a lesser extent "*Deciduous forests on moist to hydromorphic soils*" and "*Deciduous forest on lichen-covered ground*" (5 to 6 observations, in 2.3 to 2.8% of the total sum of polygons).

The χ^2 test is very significant (observed value 383.1, critical value 5.99, $p<0.0001$, Table IV), the over-representation of Birch trees in Broadleaved forests matching up with 92% of the observed χ^2 value.

TABLE IV. CHI-SQUARED VALUES (X^2) BETWEEN ECOLOGICAL HABITATS CATEGORIES AND MOUNTAIN BIRCH RECORDS COUNTED AMONG THOSE AT THE MICRO-REGIONAL SCALE (NORTHERN LAPLAND)

Ecological habitat X_2	**Mountain Birch X_2**		
	Σ absences	***Σ presences***	***Total***
Habitats without Mountain Birch tree	0.1	14	14.06
Habitats where Mountain Birch is lowly observed	0.12	16.7	16.73
Deciduous forests habitat	2.55	349.77	352.31
Total	2.77	380.33	383.1

b) Study case of Caluna AB data provider at a local scale with photointerpretation

As shown in the Figure 6, Birch records are concentrated on a narrow strip inferior to 100 m from the railway line (93.4 m on average), while photo-interpreted Birch stands are much more widespread. To the eastern part of the studied local site, the south shore of the line is very sloping, explaining the very low number of occurrences (N=2). The 11 other observations are made to the north, between the railway line and the E10 road, i.e. in an easily accessible area. Birch stands extend over more than one kilometer and a half (up to 1.7 km), but the numerous water bodies and wetlands (21% of the surface) probably constituted obstacles for observers. In the west, observations are made on the south side along a secondary road between a train station and the E10 highway, then observations are made on the north side when the south slope is rocky and partially snowy, without pedestrian access.

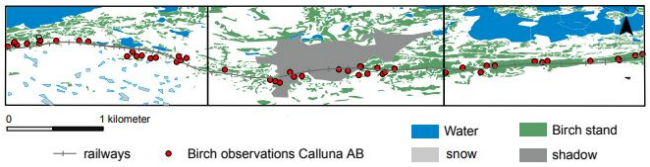

Fig. 6. Local photointerpretation displaying relationship between Birch records, transportation and Birch stands (Sources: GBIF; realization M.Cohen 2020)

The number of mountain Birch observations is not proportional to the area occupied by forest formations. Their density is 0.12 in the east and 0.29 in the west. This result strongly suggests that observations and records were more numerous while the presence of the species was more remarkable because of its scarcity within the landscape (e.g. observer-bias).

IV. DISCUSSION AND PERSPECTIVES

a) OpenAccess biodiversity databases (OBDs) spatial variation in regard of geographical laws

By previous results, a discussion about the relevance of the general laws in geography, especially in biogeography and spatial ecology has been raised.

Indeed, the principle of geographical similarity (as defined as similar configurations of two locations having similar features) has been well illustrated when looking at records distribution of Birch tree even before habitat niche modeling. Mainly, this study was dedicated to carefully analyze metadata and their ecological meaning applied to habitat niche modeling (e.g. abiotic factors describing and explaining records distribution). Geographical similarity has been evoked all along the study, in particular at the micro-regional and local scale of spatial analysis. Such similarity was geographical-dependent, mainly calling for altitudinal and topography factors, where mountain Birch woodlands oftenly are found in the shallows of glacial valleys, all along the shore of lakes, open waters and watercourses simultaneously with human facilities such as roads, railways and housings. Finally, law of similarity appears to be very useful to distinguish autocorrelation effects, either be natural or human-induced through sampling biases. Geographical laws allow to disentangle the role of sampling bias, expressed by autocorrelation law, and the role of environmental conditions, corresponding to the geographical similarity principle.

b) Applicability of the study for habitat niche modeling and perspectives

The main descriptor being latitude and longitude of mountain Birch records, analyzing links between institution providers, given by the GBIF dataset has been highly relevant. Spatial autocorrelation having compared different providing institutions has been promising in the frame of the aim: at the macro-regional scale, the distribution of Birch tree records for the Norrbotten county (Artdatabanken as providing institution) appears to be more suitable for habitat niche

modeling. To support those first conclusions, and considering the dominant administrative and institutional (public) source of records, administrative boundaries of Swedish Lapland (from province, to counties until municipalities) would be interesting to select a dataset where records are not spatially clustered, at least not because of human factors (e.g. accessibility, facility).

c) Learning for Geography

From this experience, we can preconize a reasoned use of Open Access databases for modeling in geography, regardless of the type of data. First of all, it is important to keep in mind geographic laws as a theoretical framework, even if they are not understood or known by the producers of the data and the scientists using them. It is sometimes the case for OBDs, only the law of spatial auto-correlation has been identified as an obstacle (Silero and Barbosa, 2020), whereas the principle of geographic similarity was of great importance and formed the basis for reliable modeling. Procedures for correcting sampling bias must likewise be reexamined in the light of an analysis that balances the difference between real bias and a location corresponding to a certain geographical reality, according to the principle of geographical similarity. For instance, caution should be exercised if it is necessary to select or remove self-correlated observations based on geographic objects, when these geographic objects meet the environmental conditions for the presence of the observations.

V. References

AGNAN, Yannick, COURAULT, Romain, ALEXIS, Marie A., et al. Distribution of trace and major elements in subarctic ecosystem soils: Sources and influence of vegetation. Science of the Total Environment, 2019, vol. 682, p. 650-662.

ANDERSON, Robert P., ARAÚJO, Miguel B., GUISAN, Antoine, et al. Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. Frontiers of Biogeography, 2020

ANDERSSON, Rikard, ÖSTLUND, Lars, et TÖRNLUND, Erik. The last European landscape to be colonised: a case study of land-use change in the far north of Sweden 1850-1930. Environment and History, 2005, vol. 11, no 3, p. 293-318.

BOILLAT, Sébastien et IFEJIKA SPERANZA, Chinwe. IPBES Global Assessment Report on Biodiversity and Ecosystem Services. Chapter 3. Assessing progress towards meeting major international objectives related to nature and nature's contributions to people. 2019.

BRYN, Anders et POTTHOFF, Kerstin. 20th century Betula pubescens subsp. czerepanovii tree-and forest lines in Norway. Biodiversity Data Journal, 2017, no 5.

COURAULT, Romain, DUVAL, Gregory, et COHEN, Marianne. La fragmentation des paysages de l'élevage des rennes. Une étude de cas en Laponie suédoise. Géocarrefour, 2018, vol. 92, no 92/3.

COURAULT, Romain. Les Paysages Culturels de L'élevage de Rennes en Scandinavie Face au Changement Global, Une Approche Multi-Scalaire (Laponie Suédoise, Sud Norvégien). Ph.D. Thesis, Sorbonne-Université, Paris, France, 2018.

EL‐GABBAS, Ahmed et DORMANN, Carsten F. Improved species‐occurrence predictions in data‐poor regions: using large‐scale data and bias correction with down‐weighted Poisson regression and Maxent. Ecography, 2018, vol. 41, no 7, p. 1161-1172

FIELD, Christopher B. (ed.). Assessment Report 5 Climate change 2014–Impacts, adaptation and vulnerability: Regional aspects. Cambridge University Press, 2014.

FORBES, Bruce C. et KUMPULA, Timo. The ecological role and geography of reindeer (Rangifer tarandus) in northern Eurasia. Geography Compass, 2009, vol. 3, no 4, p. 1356-1380.

GOUNOT, M. Méthodes quantitatives d'étude de la végétation, Editions Masson & Cie, Paris. 1969.

KULLMAN, Leif. Tree limit dynamics of Betula pubescens ssp. tortuosa in relation to climate variability: evidence from central Sweden. Journal of Vegetation Science, 1993, vol. 4, no 6, p. 765-772.

KULLMAN, Leif. Climate change and primary birch forest (Betula pubescens ssp. czerepanovii) succession in the treeline ecotone of the Swedish Scandes. International Journal of Research in Geography, 2016, vol. 2, no 2, p. 36-47.

KUMPULA, Jouko, STARK, Sari, et HOLAND, Øystein. Seasonal grazing effects by semi-domesticated reindeer on subarctic mountain birch forests. Polar Biology, 2011, vol. 34, no 3, p. 441-453.

MANKER, Ernst Mauritz. The nomadism of the Swedish mountain lapps: the siidas and their migratory routes in 1945. H. Geber, 1953.

PEARSON, Richard G., PHILLIPS, Steven J., LORANTY, Michael M., et al. Shifts in Arctic vegetation and associated feedbacks under climate change. Nature climate change, 2013, vol. 3, no 7, p. 673-677.

PFEIFER, Susanne, RECHID, Diana, REUTER, Maximilian, et al. 1.5, 2, and 3 global warming: visualizing European regions affected by multiple changes. Regional Environmental Change, 2019, vol. 19, no 6, p. 1777-1786.

RUETE, Alejandro. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. Biodiversity Data Journal, 2015, no 3.

RYDIN, Håkan, SNOEIJS, Pauli, et DIEKMANN, Martin. Swedish plant geography: dedicated to Eddy van der Maarel on his 65th birthday. Svenska växtgeografiska sällsk., 1999.

SEGURADO, P. A. G. E., ARAÚJO, Miguel B., et KUNIN, W. E. Consequences of spatial autocorrelation for niche‐based models. Journal of Applied Ecology, 2006, vol. 43, no 3, p. 433-444.

SILLERO, Neftalí et BARBOSA, A. Márcia. Common mistakes in ecological niche models. International Journal of Geographical Information Science, 2020, p. 1-14

SIZYKH, Alexander P., SIZYKH, Svetlana V., et al. Examples of Ecotones and Paragenese in the Vegetation Cover of the Baikalian Siberia. Natural Science, 2014, vol. 6, no 15, p. 1197.

SVONNI, Ragnhild. Samisk markanvändning och MKB. Sametinget, 2010.

TRUONG, Camille, PALMÉ, Anna E., et FELBER, François. Recent invasion of the mountain birch Betula pubescens ssp. tortuosa above the treeline due to climate change: genetic and ecological study in northern Sweden. Journal of evolutionary biology, 2007, vol. 20, no 1, p. 369-380.

TUTIN, Thomas Gaskell, HEYWOOD, Vernon Hilton, BURGES, Norman Alan, et al. (ed.). Flora Europaea: Plantaginaceae to Compositae (and Rubiaceae). Cambridge University Press, 1964.