Katedra informatiky
Přírodovědecká fakulta
Univerzita Palackého v Olomouci

DIPLOMOVÁ PRÁCE

Extrakce a analýza dat z EMMA databáze

2019

Bc. Miroslav Michálek

Vedoucí práce:
Mgr. Martin Trnečka, Ph.D.

Studijní obor: Aplikovaná informatika,
prezenční forma

Bibliografické údaje

Autor:	Bc. Miroslav Michálek
Název práce:	Extrakce a analýza dat z EMMA databáze
Typ práce:	diplomová práce
Pracoviště:	Katedra informatiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci
Rok obhajoby:	2019
Studijní obor:	Aplikovaná informatika, prezenční forma
Vedoucí práce:	Mgr. Martin Trnečka, Ph.D.
Počet stran:	52
Přílohy:	1 CD/DVD
Jazyk práce:	český

Bibliograhic info

Author:	Bc. Miroslav Michálek
Title:	Data analysis from the EMMA database
Thesis type:	master thesis
Department:	Department of Computer Science, Faculty of Science, Palacký University Olomouc
Year of defense:	2019
Study field:	Applied Computer Science, full-time form
Supervisor:	Mgr. Martin Trnečka, Ph.D.
Page count:	52
Supplements:	1 CD/DVD
Thesis language:	Czech

Anotace

Diplomová práce se zabývá extrakcí, následným sloučením a analýzou dat ze dvou rozdílných databází. Prvním cílem této práce je extrakce dat z databáze evropských savců a dat světové klimatické databáze. Druhým cílem je sloučit data z obou databází do datasetu ve formě Booleovské matice a dále data analyzovat za pomocí metody rozkladu matic.

Synopsis

This thesis deals with the extraction, the following merge and the analysis of the data from two different databases. The first aim of the study is to extract the data from the european mammal database and the world climatic database. The second aim of the study is to merge the data from both databases as a Boolean matrix and to analyse the data using the matrix decomposition method.

Klíčová slova: analýza faktorových konceptů; extrakce dat; EMMA; WorldClim;

Keywords: analysis of factor concepts; data extraction; EMMA; WorldClim;

Děkuji Mgr. Martinu Trnečkovi, Ph.D. za cenné rady a odborné vedení práce. Dále bych rád poděkoval své rodině.

Místopřísežně prohlašuji, že jsem celou práci včetně příloh vypracoval samostatně a za použití pouze zdrojů citovaných v textu práce a uvedených v seznamu literatury.

datum odevzdání práce podpis autora

Obsah

Seznam obrázků

Seznam tabulek

Seznam vět

Seznam zdrojových kódů

1 Úvod

Diplomová práce řeší problém extrakce a následné analýzy dat ze dvou rozdílných databází. Hlavním cílem práce je vytvoření datasetu obsahujícího informace o výskytu evropských savců spolu s informacemi o klimatických podmínkách panujících v Evropě. Dalším cílem práce je pomocí moderní metody rozkladu matic získat netriviální a potenciálně zajímavé informace z datasetu.

Práce je rozdělena do tří části. V první části je představen postup extrakce dat ze dvou databází včetně sloučení dat do jednoho datasetu ve formě Booleovské matice. Druhá část je věnována moderní metodě rozkladu matic - GreConD. Poslední část práce obsahuje interpretaci a vizualizaci nalezených informací.

1.1 Vysvětlení základní pojmů

Vysvělení základních pojmů vyskytujích se v textu této práce.

1.1.1 XML

e**X**tensible **M**arkup **L**anguage je značkovací jazyk, vyvinut a standardizován konsorciem W3C. Navržen pro uchovávání a transport dat. XML nevyužívá předdefinované tagy, narozdíl od HTML. Autor XML dokumentu vytváří vlastní tagy. Rovněž autor definuje strukturu dokumentu.

1.1.2 SVG

Scalable **V**ector **G**raphics (škálovatelná vektorová grafika) je rozšířený a bezplatný grafický formát vyvíjen pracovní skupinou W3C. Popisuje vektorovou grafiku pomocí XML značek. Základní otevřený formát pro vektorovou grafiku na webových stránkách, od HTML5 mohou být SVG elementy vkládány přímo do HTML stránek.

1.1.3 TIFF

TIFF (z anglického Tagged Image File Format) je formát pro rastrovou grafiku. Vytvořen společností Aldus v roce 1986. Umožňuje ukládat obrazová data v komprimované i nekomprimované podobě. Adaptabilita formátu je zaručena díky možnosti přídávání vlastních tagů do záhlaví souboru. Tagy mohou obsahovat další informace vztahující se k datům v obrazové části.

2 Extrakce dat z EMMA databáze

2.1 EMMA

Databáze evropských savců (dostupná na http://www.european-mammals.org). Zde dostupná data vychází z knihy [1]. Databáze obsahuje informace o výskytu evropských savců.

2.2 Popis extrakce

Během tvorby této práce nebylo dostupné API pro pohodlné získání dat z databáze, data byla získávána postupně přímo z HTML struktury webové stránky.

Seznam všech dostupných savců je vytažen z elementu `<select>`, přesněji ze všech možností <option>. Viz zdrojový kód 1. Celkem je dostupných 245 savců na území celého evropského kontinentu. Rodová a druhová jména savců jsou uváděna v latině.

```
1 <select name = "latname" size="1">
2   <option>Macropus rufogriseus</option>
3   <option>Atelerix algirus</option>
4   <option>Erinaceus concolor</option>
5 </select>
```

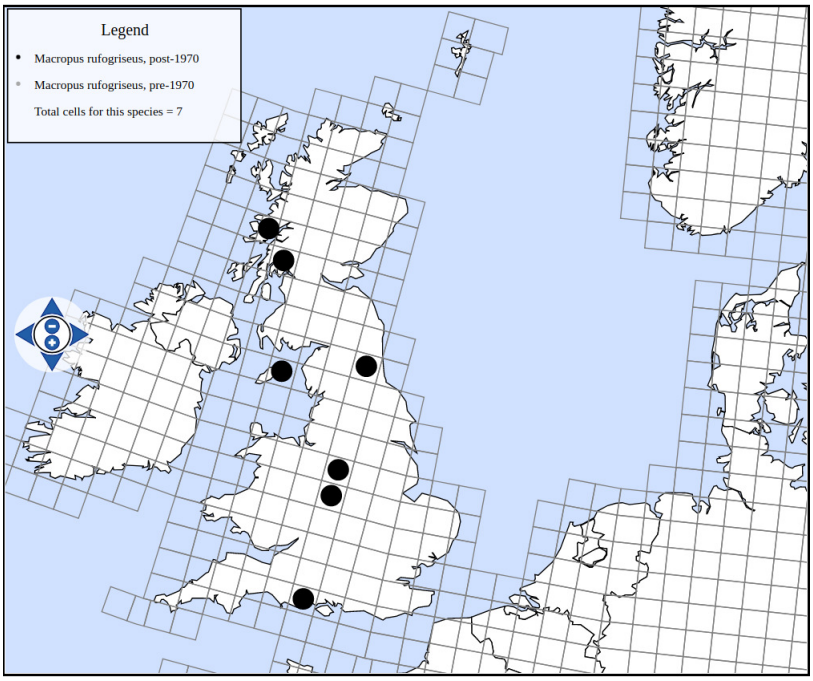
Zdrojový kód 1: Ukázka části select boxu

Program pro extrakci využívá vyhledávání dat v režimu „Atlas Style Mapping“. Toto vyhledávání je dostupné na webových stránkách databáze evropských savců. Po zvolení hledaného savce je vrácena mapa Evropy ve formátu SVG. Pevninskou část mapy překrývá čtvercová mřížka, výskyt jednotlivého savce je zobrazen ve formě černých a šedých bodů vyplňujících příslušná políčka mřížky, viz kód 2. Šedé body symbolizují výskyt savce před rokem 1970, černé po roce 1970. Pro účely práce jsou brány v potaz pouze výskyty po roce 1970, čili černé body. Každé políčko mřížky má svůj unikátní CGRS identifikátor. CGRS identifikátor je alfanumerický řetězec znaků ukazující na buňku CGRS mřížky, viz obrázek 2.

```
1 <circle xmlns="http://www.w3.org/2000/svg" id="post70_1" r="200"
    fill="black"/>
```

Zdrojový kód 2: SVG element reprezentující výskyt savce

Body ve čtvercové mřížce reprezentující výskyt klokana rudokrkého (*Macropus rufogriseus*) na uzemí Anglie, viz obrázek 1.



Obrázek 1: Klokan rudokrký v Anglii

Pro zjištění CGRS identifikátorů zaplněných buněk se zaklesne do struktury SVG obrázku. Po nalezení elementu s `id=„datapoints“` stačí získat id všech jeho potomků. Tímto způsobem dojde k získání lokací výskytu jednotlivých savců, respektive k získání identifikátorů buněk mřížky. Ukázka hledaného elementu viz zdrojový kód 3.

```
<g xmlns="http://www.w3.org/2000/svg"
  id="datapoints"
  style="pointer-events:all"
  onmouseover="currentcoord(evt)"
  onmouseout="hideutm(evt)">
  <use id="30UVF2" x="28271.03" y="31441.26"
    xmlns:xlink="http://www.w3.org/1999/xlink"
      xlink:href="#post70_1"/>
  <use id="30UWB4" x="28673.33" y="35672.53"
    xmlns:xlink="http://www.w3.org/1999/xlink"
      xlink:href="#post70_1"/>
</g>
```

Zdrojový kód 3: Element s lokacemi savců

Pro přesné určení polohy je k dispozici JSON soubor EMMAGrid. Soubor zařazuje výše zmíněnou mřížku do kontextu reálného světa. Jednotlivé čtyřúhelníky mřížky propojuje s GPS souřadnicemi pomocí CGRS identifikátorů. Ve viditelné části webu není soubor k nalezení, objevit jej lze pomocí vývojářských nástrojů, které nabízí prohlížeče. Dostupný zde: https://www.european-mammals.org/osm/EMMA2grid.php. Ukázka části souboru viz 4.

```
{"type":"Feature","properties":{"CGRSName":"id"},
"geometry":{"type":"Polygon",
"coordinates":[[[long1,lat1]...[long5,lat5]]]}}
```

Zdrojový kód 4: Ukázka reprezentace mřížky

Po úspěšné extrakci jsou informace z EMMA databáze uloženy na disk. Při opětovném běhu programu není potřeba procházet webovou stránku. Využijí se následující soubory:

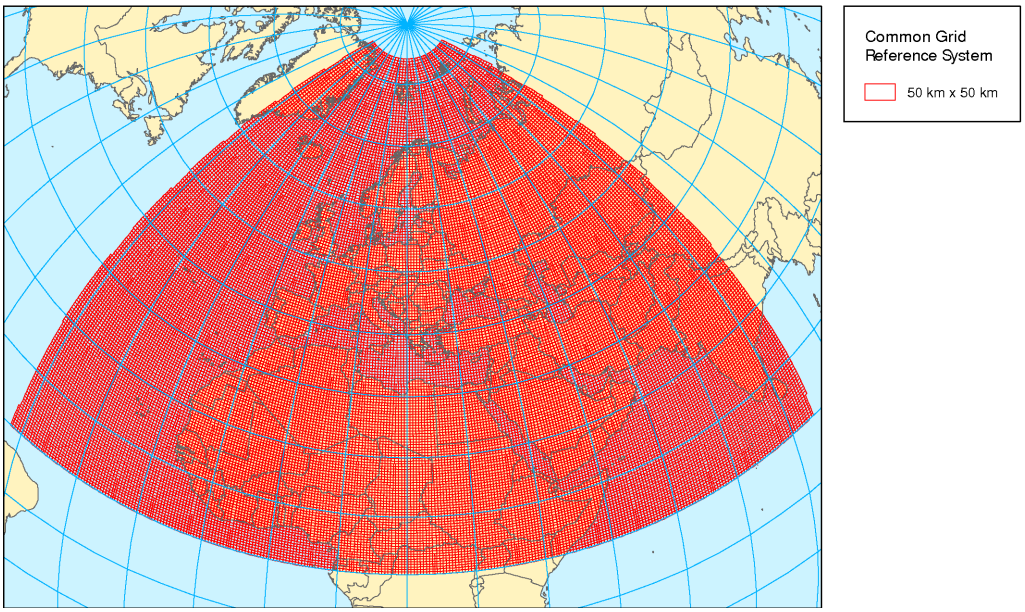
- `mammals.xml` - XML soubor obsahující informace o savcích, jméno savce a lokace výskytu. Část tohoto souboru je zobrazena viz zdrojový kód 5.
- `CGRSJSON` - soubor s GPS koordináty pro lokace výskytu.

```
<Mammal>
  <name>Sorex+isodon</name>
  <preLocations>
    <preLocation>33VUH3</preLocation>
  </preLocations>
  <posLocations>
    <posLocation>32VPP2</posLocation>
    <posLocation>33VUJ4</posLocation>
    <posLocation>33VWK3</posLocation>
    <posLocation>34VFN1</posLocation>
  </posLocations>
</Mammal>
```

Zdrojový kód 5: Ukázka ze souboru mammals.xml

2.2.1 CGRS

CGRS (z anglického Common European Chorological Grid Reference System) byl vytvořen v roce 2000 zástupci atlasových skupin mapujících evropské savce, ptáky, obojživelníky, plazy, bezobratlé, cévnaté rostliny a houby. Užíván pro mapování distribuce jednotlivých druhů. Vychází z MGRS (z anglického Military Grid Reference System) což je alfanumerická verze numerické souřadnicové soustavy UTM (z anglického Universal Transverse Mercator). Kompletní mřížka zobrazena na obrázku 2.



Obrázek 2: Kompletní CGRS mřížka

3 Extrakce dat z WorldClim databáze

3.1 WorldClim

WordClim databáze [2] nabízí zdarma k dispozici klimatická data z období 1970 – 2000. Jak název napovídá, databáze sleduje klimatické proměnné z celého světa, sledovované proměnné:

- minimalní teplota (°C)
- maximální teplota (°C)
- průměrná teplota (°C)
- srážky (mm)
- sluneční záření (kJ m^{-2} day^{-1})
- rychlost větru (m/s^{-1})
- tlak vodní páry (kPa)

Data je možno stáhnout ve formátu rastrového obrázku GeoTIFF. K dispozici jsou čtyři prostorové rozlišení. Od velikosti ∼340 km^2 až po ∼1 km^2 na jednu sledovovanou oblast. S hustotou rozlišení ovšem rapidně roste velikost souborů, pro účely práce je tedy použita varianta nejnižšího rozlišení. Pixely pokrývají sledovované oblasti a jejich hodnoty. Hodnoty pixelů odpovídají hodnotám proměnných v oblastech, naměřených za sledovované období. Při velikosti ∼340 km^2 na oblast má rastrový obrázek velikost 2160×1080 pixelů. Data jsou ve formátu single-precision 32-bit IEEE 754 floating point, neboli v pohyblivé řádové čárce se základní přesností.

Pro každou proměnnou existuje 12 souborů ve formátu GeoTIFF. Jednotlivé měsíce v roce jsou zastoupeny vlastním souborem. Soubor obsahuje zprůměrovaná data naměřená v daném měsíci za celé sledovované období třiceti let při vybrané hustotě rozlišení. Při prvním spuštění dojde ke stažení archivů z webové stránky WorldClim, následné extrahování archivů a uložení do kořenové složky pro další použití.

3.2 GeoTIFF

Specifikace GeoTIFF poskytuje množinu TIFF tagů pro popsání kartografických informací uložených ve formě rastrové grafiky, která byla pořízena satelitním nebo leteckým snímkováním, či dalšími způsoby. Tato specifikace umožňuje vázání rastrového obrázku na známá mapová zobrazení a zároveň tato zobrazení popisuje. Geotagy obrázku z databáze WorldClim viz 6.

```
Geotiff_Information:
  Version: 1
  Key_Revision: 1.0
  Tagged_Information:
    ModelTiepointTag (2,3):
      0            0            0
      -180         90           0
    ModelPixelScaleTag (1,3):
      0.166666666666667 0.166666666666667 0
    End_Of_Tags.
  Keyed_Information:
    GTModelTypeGeoKey (Short,1): ModelTypeGeographic
    GTRasterTypeGeoKey (Short,1): RasterPixelIsArea
    GeographicTypeGeoKey (Short,1): GCS_WGS_84
    GeogCitationGeoKey (Ascii,7): "WGS 84"
    GeogAngularUnitsGeoKey (Short,1): Angular_Degree
    GeogSemiMajorAxisGeoKey (Double,1): 6378137
    GeogInvFlatteningGeoKey (Double,1): 298.257223563
    End_Of_Keys.
  End_Of_Geotiff.
```

Zdrojový kód 6: Geotagy tiffu WorldClim

ModelTiepointTag $= (..., i, j, k, x, y, z...)$ váže zeměpisný souřadnicový systém s pixelovým systémem souřadnic. Pixel (i, j) leží na souřadnicích (x, y) cílového modelu. Konkrétně tedy pixel $(0, 0)$ se nachází na 180° západní délky a 90° severní šířky. Oba systémy jsou dvou-dimenzionální, tudíž $k, z = 0$.

ModelPixelScaleTag $= (scaleX, scaleY, scaleZ)$ určuje vzdálenost pixelů v jednotkách cílového modelu. Parametry $scaleX$ a $scaleY$ určují horizontální a vertikální vzdálenost rastrových pixelů.

Uvedené dva tagy dohromady určují vztah mezi rastrovým a modelovým prostorem, v tomto případě prostorem zeměpisných souřadnic. Převod do prostoru zeměpisných souřadnic:

$$Longitude = ModelTiePointTag(x) + pixel(i) * scaleX$$
$$Latitude = ModelTiePointTag(y) + pixel(j) * scaleY$$

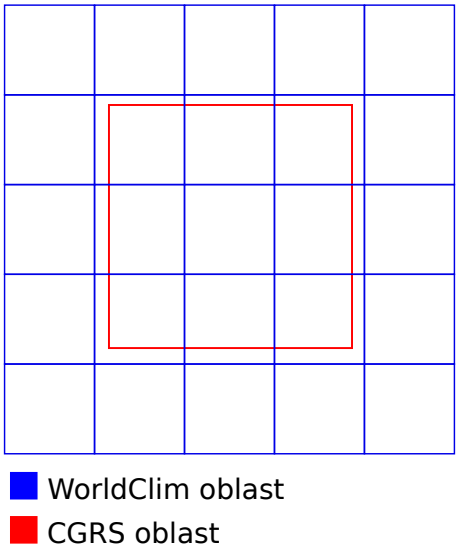
4 Tvorba matice

Pro potřeby následné analýzy byla extrahovaná data sloučena do jednoho datasetu, který je ve tvaru Booleovské matice. Množina řádků (objektů), v konkrétním případě savců, čítá 245 členů. Množina sloupců (vlastností objektů) má 395 760 členů. Prvních 4656 sloupců (vlastností) patří oblastem CGRS mřížky. Následující sloupce náleží hodnotám klimatických proměnných.

4.1 Sloučení datasetů

Vytvoření výsledného datasetu vyžadovalo přiřadit hodnoty klimatických proměnných k oblastem CGRS mřížky, se kterými pracuje evropská databáze savců. Přiřazení oblastí výskytu savců do oblastí z klimatické databáze bylo uvažováno, nicméně z důvodu již stanovených unikátních identifikátorů a koordinátů u CGRS mřížky byla vybrána první varianta.

Oblasti CGRS mřížky mají až na vyjímky rozměry 50 km × 50 km, tedy přibližně 2500 km^2. Sledovované oblasti WorldClim databáze mají plochu $\sim$ 340 km^2.



Obrázek 3: Mřížky

Prvním krokem pro výpočet hodnoty klimatické proměnné v CGRS oblasti bylo určení středových koordinátů. V programu je oblast reprezentována objektem třídy **Area**. Objekt obsahuje jediný slot GPSCoordinates, což je pole obsahující koordináty vymezující danou oblast. K výpočtu středu slouží metoda **center()**. Výsledná hodnota je vypočtena pomocí vzorce pro výpočet středu polygonu.

Ve druhém kroku se vyhledává v datech klimatické databáze. Pro získaný bod se středovými koordináty je nalezen pixel, který tento bod obsahuje. Zde je zavolána metoda **coorsToPixel()** viz kód 7. Metoda používá GeoTIFF tagy **ModelTiepointTag** a **ModelPixelScaleTag** zmíněné na předchozí straně.

```
1 public int[] coorsToPixel(Location loc) {
2 return new int[]{(int) ((loc.getLongitude() - lonStart) / xScale),
3      (int) ((loc.getLatitude() - latStart) / yScale) * -1}; }
```

Zdrojový kód 7: Metoda coorsToPixel

Nyní je známa poloha pixelu, který obsahuje střed větší CGRS oblasti z databáze savců, následně je třeba zvolit dostatečně velké okolí pixelu k pokrytí celé této oblasti. Přistupuje se ke všem pixelům do vzdálenosti dva, tedy 24-okolí.

Pro všech 25 pixelů ze vzniklé masky se vypočítá procentuální překrytí s oblastí CGRS. K dispozici jsou prozatím pouze souřadnice pixelů, nikoliv jejich přesná poloha na Zemi. Funkcionalitu zjistění polohy vymezujících bodů pixelu zajišťuje metoda **getPixelBoundaries()**, ve svém těle zavolá na vrcholy čtyřúhelníku další metodu **getPxLocation()** a dojde k vypočtení přesných souřadnic bodů vymezující oblast, výpočet viz zdrojový kód 8.

```
1 coors[0] = lonStart + x * xScale;
2 coors[1] = latStart - y * yScale;
```

Zdrojový kód 8: Výpočet přesné polohy

Jedná se o reverzi výpočtu uvedeného v metodě **coorsToPixel()**. Po aplikaci na všechny pixely pod maskou lze vypočíst plochu, kterou tyto pixely reprezentují. U referenční buňky CGRS rovněž vypočítáme plochu, přesné hraniční souřadnice jsou již známy. Samotný výpočet plochy obstarává metoda **getArea()**, implementuje výpočet plochy konvexního mnohoúhelníku [3].

$$A = \frac{1}{2}(x_1y_2 - x_2y_1 + x_2y_3 - x_3y_2 + \cdots + x_{n-1}y_n - x_ny_{n-1} + x_ny_1 - x_1y_n)$$

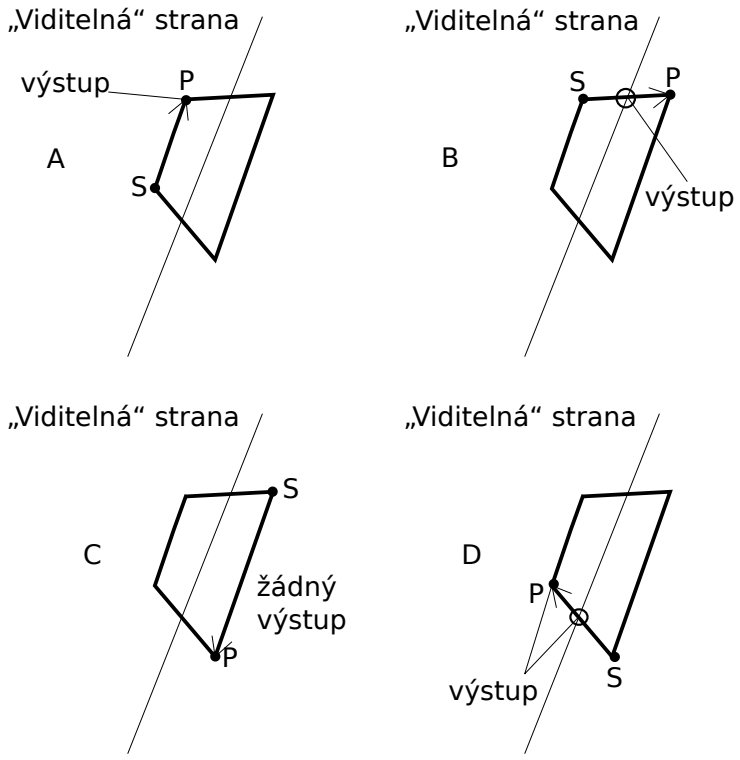
V programu jsou souřadnice vrcholů čtyřúhelníku uloženy proti směru hodinových ručiček. Před návratem z metody je nutno vypočítat vypočítat absolutní hodnotu.

4.1.1 Výpočet průniku oblastí

V této fázi je již možné vypočítat procentuální překrytí jednotlivých oblastí masky s referenční oblastí. Předně je nutné nalézt průnik oblastí. Vlastní metoda třídy **Area intersection(Area area)** vrací průnik dvou daných oblastí ve formě objektu třídy **Area**.

Pro nalezení průniku dvou mnohoúhelníků, je v práci použit algoritmus **Reentrant Polygon Clipping** [4] ve verzi pro dvě dimenze. Algoritmus ořezává polygony vůči irregulárním konvexním oknům. Výsledné polygony jsou vraceny

ve formě uspořádané posloupnosti vrcholů, tedy v konvenci, která je již v programu zavedena. Polygon je postupně ořezáván vůči každé ořezové rovině, či hranici ořezového okna.
Mnohoúhelník je definován jako uspořádaná množina vrcholů $P_1, \cdots, P_n$, cílem je najít novou množinu $Q_1, \cdots, Q_n$ bodů nacházejících se na „viditelné“ straně ořezové roviny, tato množina reprezentuje nově vzniklý mnohoúhelník. Algoritmus při ořezávání vůči jedné rovině navštíví každý bod z $P_1, \cdots, P_n$ právě jednou. Při každém navštívení P_i je vygenerováno 0–2 bodů náležících $Q_1, \cdots, Q_n$, znázorněno viz obrázek 4.



Obrázek 4: Ořezové pravidlo

Vztah mezi hranou a ořezovou rovinou, hranou ořezového okna je charakterizován čtyřmi možnými případy:

- V případě A, kde se celá hrana, tedy již zpracovaný vrchol S a zpracovávaný vrchol P nachází na „viditelné“ straně, je výstupem bod P, bod S byl zpracován v předchozím kroku.
- B znázorňuje případ, kdy hrana vystupuje z „viditelné“ strany, bod S se nachází uvnitř, bod P je mimo. Zde na výstup přijde průsečík hrany a ořezové roviny.
- Ve variantě C je celá zpracovávaná hrana mimo „viditelnou“ stranu, tudíž na výstup nejde žádný z bodů.

- V případě D hrana vstupuje na „viditelnou“ stranu, výstupem je zpracovávaný bod P a průsečík hrany a roviny.

Pro rozpoznání možného vztahu mezi hranou a ořezovou rovinou bylo třeba zavést metodu **boolean isFront(double[] line, double[] point)** implementující vzorec pro vzdálenost bodu od úsečky.

$$distance(P_1, P_2, (x_0, y_0) = \frac{(y_2 - y_1)x_0 - (x_2 - x_1)y_0 + x_2y_1 - y_2x_1}{\sqrt[2]{(y_2 - y_1)^2 + (x_2 - x_1)^2}}$$

Pokud je výsledná hodnota ≥ 0, vrací metoda pravdivostní hodnotu **true**, jinak **false**.

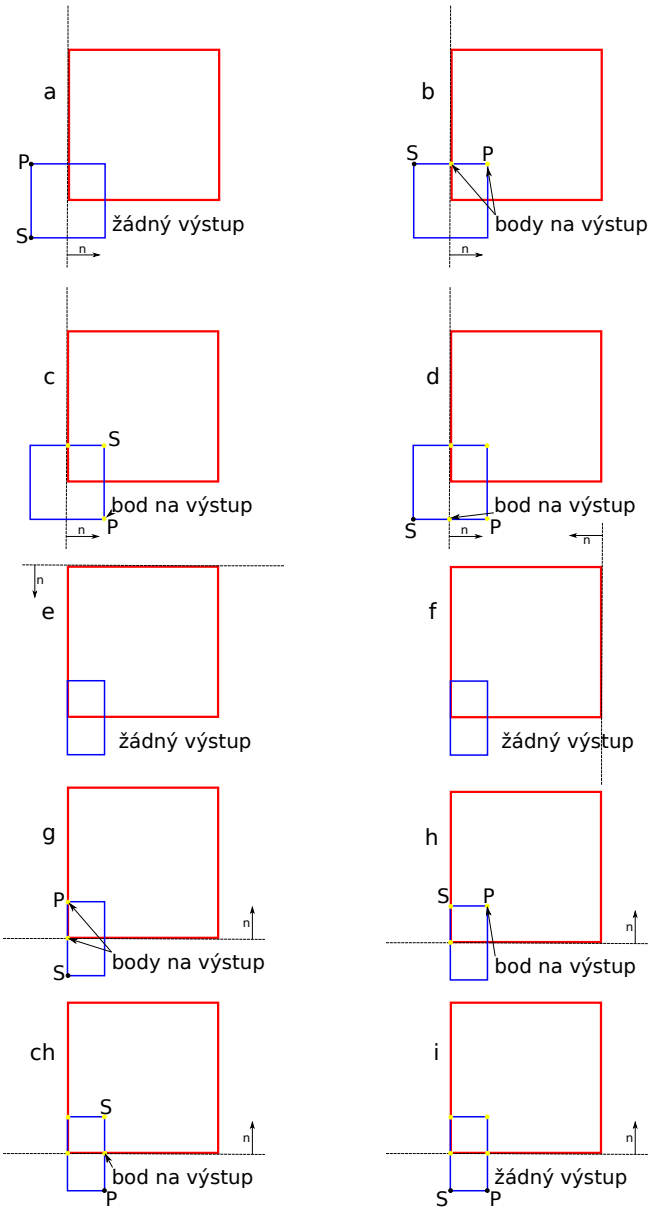
Výpočet souřadnic bodu průniku, znázorněného ve případech 4B a 4D, zajišťuje metoda **double[] intersection(double[] line, double[] p, double[] q)**, implementující následující vzorec [5]

$$x = \frac{\begin{vmatrix} \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} & x_1 - x_2 \\ \begin{vmatrix} x_3 & y_3 \\ x_4 & y_4 \end{vmatrix} & x_3 - x_4 \end{vmatrix}}{\begin{vmatrix} x_1 - x_2 & y_1 - y_2 \\ x_3 - x_4 & y_3 - y_4 \end{vmatrix}}$$

$$y = \frac{\begin{vmatrix} \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} & y_1 - y_2 \\ \begin{vmatrix} x_3 & y_3 \\ x_4 & y_4 \end{vmatrix} & y_3 - y_4 \end{vmatrix}}{\begin{vmatrix} x_1 - x_2 & y_1 - y_2 \\ x_3 - x_4 & y_3 - y_4 \end{vmatrix}}$$

Poslední metodou, využívanou při hledání oblasti průniku, je metoda **List <double[]> getLines(double[] polygon)**, která vrací pro polygon z parametru seznam jeho hran, jednotlivá hrana je ve formátu (x_1, y_1, x_2, y_2).

Prozatím bylo ukázáno ořezávání mnohoúhelníku vůči jedné rovině. Při sjednocování datasetů byl ale vždy ořezáván menší čtyřúhelník oblasti WorldClim, vůči větší oblasti CGRS (ořezové okno). Jak lze pozorovat na příkladu, ořezání vůči oknu, je pouhá posloupnost ořezání vůči rovinám, které tvoří hranice okna. V první iteraci je vybrána levá hrana okna, tedy první bod v poli bodů reprezentují okno a bod následující. Provede se uplatnění pravidel ořezu na hrany ořezávaného čtyřúhelníku a výsledné body jsou předány do další iterace. Následující iterace opět uplatní pravidla na všechny hrany mnohoúhelníku z předchozí iterace. Po navštívení všech hran okna algoritmus končí.



Obrázek 5: Postupné ořezávání

V první iteraci, případech a–b z obrázku 5 dojde k uplatnění výše uvedených pravidel a do další iterace je předán čtyřúhelník vyznačen žlutými body.

Druhá a třetí iterace e–f není detailně předvedena, protože pro každou stranu ořezávaného čtyřúhelníku je uplatněno první pravidlo, tedy body S, P se nachází na „viditelné“ straně a bod P jde na výstup. V obou případech je předán čtyřúhelník nezměněn.

Poslední iterace g–i vrátí finální ořezaný čtyřúhelník (žluté body) odpovídající průniku zobrazených.

4.1.2 Výpočet klimatické hodnoty

Pro dokončení výpočtu hodnoty klimatické proměnné na referenční oblasti je užito váženého průměru z již představeného okolí. Soubor n hodnot

$$X = \{x_1, \cdots, x_n\},$$

kde jsou zastoupeny hodnoty klimatických proměnných naměřených v oblastech, které reprezentují pixely GeoTiff obrázku WorldClim. Dále jsou k dispozici odpovídající váhy

$$W = \{w_1, \cdots, w_n\},$$

rovnající se procentuálnímu překrytí jednotlivých klimatických oblastí s referenční oblastí. Překrytí počítano viz zdrojový kód 9.

```
1 coverages[k] = refPoly.intersection(getPixelBoundaries(i,j))
2     .getArea();
3 coverages[k] = coverages[k] / refPolyArea * 100;
```

Zdrojový kód 9: Výpočet překrytí

Pak vážený průměr je dán vzorcem

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

Po aplikaci všech zmíněných kroků je získána hodnota klimatické proměnné pro zpracovávanou oblast. Celý postup se opakuje na všech buňkách CRGS mřížky pro všechny sledovované klimatické proměnné ve všech sledovovaných obdobích. Postupně se přistoupí k 84 staženým obrázkům databáze WorldClim.

Všechny kroky potřebné před samotnou tvorbou matice jsou vykonány, lze tedy přejít k vytvoření matice. Matice má 245 řádků, odpovídá počtu získaných savců. Sloupců má matice násobně více, celkový počet se zastaví na čísle 395 760.

Prvních 4656 sloupců reprezentuje oblasti ve kterých se savec může vyskytovat. Pokud některý z prvků $a_{n,i}$, kde $i \leq 4656$ je roven 1, pak savec reprezentovaný řádkem n se vyskytuje na oblasti reprezentované sloupcem i.

Zbylých $7 \times 12 \times 4656 = 391104$ sloupců reprezentuje hodnoty klimatických proměnných u výše zmíněných oblastí výskytu.

Pro objasnění, vyskytuje-li se savec na řádku $n = 1$ v oblasti reprezentovanou sloupcem $i = 1$, pak hodnotu první klimatické proměnné v lednu představuje sloupec s indexem $j = i + 4656$. Opětovným přičtením získáme index sloupce pro únor první klimatické proměnné. Viz obrázek 6.

$$
\begin{array}{l}
\\
\text{Savci}
\end{array}
\overbrace{\phantom{\begin{array}{ccc} a_{1,1} & \cdots & a_{1,4656} \end{array}}}^{\text{Oblasti}}
\overbrace{\phantom{\begin{array}{ccc} a_{1,4657} & \cdots & a_{1,9312} \end{array}}}^{\text{Leden 1. proměnná}}
\overbrace{\phantom{\begin{array}{ccc} a_{1,9313} & \cdots & a_{1,13968} \end{array}}}^{\text{Únor 1. proměnná}}
\left(\begin{array}{ccccccccccc}
a_{1,1} & \cdots & a_{1,4656} & a_{1,4657} & \cdots & a_{1,9312} & a_{1,9313} & \cdots & a_{1,13968} & \cdots \\
a_{2,1} & \cdots & a_{2,4656} & a_{2,4657} & \cdots & a_{2,9312} & a_{2,9313} & \cdots & a_{2,13968} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
a_{m,1} & \cdots & a_{m,4656} & a_{m,4657} & \cdots & a_{m,9312} & a_{m,9313} & \cdots & a_{m,13968} & \cdots
\end{array}\right)
$$

Obrázek 6: Popis výsledné matice

Na disku je matice reprezentována pomocí třech souborů:

- `colsHeaders.txt` - obsahuje záhlaví sloupců, CGRS identifikátory oblastí, následované vypočtenými klim. hodnotami v těchto oblastech.
- `rowsHeaders.txt` - obsahuje záhlaví řádků, latinská jména savců
- `data.txt` - data uložena v řídkém formátu, uloženy pouze indexy sloupců, které mají na průsečíku s aktuálním řádkem hodnotu *Pravda*. Při přechodu na další zpracovávaný řádek, zapíše program do souboru speciální symbol.

Záhlaví sloupců uloženo ve formátu UTF-8, v `data.txt` jsou pak jednotlivá čísla sloupců kódována do 4 bytů.

4.1.3 Řídká matice

Matice je řídká, pokud má většinu prvků nenulových nebo je uložena v řídkém formátu, tedy ukládají se pouze nenulové prvky. Zaznamenáváním nulových prvků by docházelo ke zbytečnému plýtvání místem na disku.

Uložení výsledné matice v řídkém formátu bylo nasnadě, nenulových prvků je pouhých 6,85 %. Pokud by zmíněná matice byla uložena v hustém formátu, zabírala by $245 \times 395760 = 96\,961\,200\,\text{B}$, což je přibližně 96,96 MB. V řídkém formátu zabírá matice na disku pouze $(6644535 + 245) \times 4 = 26{,}57\,\text{MB}$, úspora místa je tedy značná.

5 Formální konceptuální analýza

5.1 Neformální úvod

Formální konceptuální analýza pracuje s tabulkovými daty. V reálném světě lze identifikovat různě *objekty* kolem sebe (pes, ryba, člověk), rovněž lze těmto objektům přisoudit rozmanité *atributy* (žije na souši, žije ve vodě). Základní vztah mezi objekty a atributy pak nabývá dvou stavů, daný objekt má/nemá daný atribut.

V tabulce řádky reprezentují objekty, sloupce pak atributy, pokud objekt má daný atribut, na průsečíku tohoto řádku a sloupce je znázorněn křížek. Pro ilustraci viz tabulka 1, na řádcích lze nalézt objekty (zde rostliny a živočichové), sloupce obsahují atributy, které tyto objekty mohou vlastnit, nabývat tak vlastností za těmito atributy. Například řádek reprezentující pijavici, má křížek ve sloupci *a*, *b* a *g*, z legendy tabulky lze následně vyčíst o jaké vlastnosti se jedná. Pijavice má tedy vlastnosti *a* „potřebuje vodu k žití“ spolu s *b* „žije ve vodě“ a *g* „může se pohybovat“, naopak vlastnostmi *c* „žije na souši“, *h* „má končetiny“ a dalšími nedisponuje.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
Bodlák	×	×		×		×			
Cejn	×	×					×	×	
Fazole	×		×	×	×				
Kukuřice	×		×	×		×			
Pes	×		×				×	×	×
Pijavice	×	×					×		
Rákos	×	×	×	×		×			
Žába	×	×	×				×	×	

Tabulka 1: Formální kontext [6]

a) potřebuje vodu k žití, *b)* žije ve vodě, *c)* žije na souši, *d)* potřebuje chlorofyl k žití, *e)* má dva děložní lístky, *f)* má jeden děložní lístek, *g)* může se pohybovat, *h)* má končetiny, *ch)* kojí své potomky

Uvedená tabulka se nazývá *kontext*. Cílem formální konceptuální analýzy (FCA) je nacházet v tomto kontextu (vstupních datech) netriviální informace, tyto netriviální informace si lze představit jako shluky dat v tabulce.

Takový shluk může představovat množina objektů {Bodlák, Cejn, Pijavice, Rákos, Žába} a množina atributů {a,b}, zobrazen viz tabulka 2.

Zmíněné objekty jsou maximální možné objekty sdílející vlastnosti *a* „potřebuje vodu k žití“ a *b* „žije ve vodě“ a přitom žádné další vlastnosti již všichni společné nemají. A zároveň atributy {a,b} jsou maximální možné atributy společné objektům {Bodlák, Cejn, Pijavice, Rákos, Žába}. Žádným dalším objektům nejsou tyto atributy dohromady společné.

Tyto dvě množiny objektů a atributů se nazývají *koncept*, množina objektů v konceptu se nazývá *extent* a množina atributů se nazývá *intent*. Koncept je zvýrazněn v tabulce 2.

	a	b	c	d	e	f	g	h	i
Bodlák	×	×		×		×			
Cejn	×	×					×	×	
Fazole	×		×	×	×				
Kukuřice	×		×	×		×			
Pes	×		×				×	×	×
Pijavice	×	×					×		
Rákos	×	×	×	×		×			
Žába	×	×	×				×	×	

Tabulka 2: Formální koncept

a) potřebuje vodu k žití, *b)* žije ve vodě, *c)* žije na souši, *d)* potřebuje chlorofyl k žití, *e)* má dva děložní lístky, *f)* má jeden děložní lístek, *g)* může se pohybovat, *h)* má končetiny, *ch)* kojí své potomky

Koncept lze také interpretovat jako maximální možný obdélník nad vybranými řádky a sloupci. I když v tabulce 2 je obdélník „rozdělen“ na dvě části, pouhým přeuspořádáním řádků, by došlo ke „spojení“.

5.2 Formální úvod

Definice 1 (Formální kontext)
Formální kontext je trojice $\langle X, Y, I\rangle$, kde X a Y jsou neprázdné množiny a I je binární relace mezi X a Y, tj. $I \subseteq X \times Y$.

Prvky množiny X se nazývají objekty, prvky množiny Y se nazývají atributy. Pokud $\langle x, y\rangle \in I$, pak lze říct, že objekt x má atribut y. V předchozí podkapitole byla ukázána reprezentace formálního kontextu pomocí tabulky. Uvedená tabulka o n řádcích a m sloupcích odpovídá formálnímu kontextu $\langle X, Y, I\rangle$ složeného z množiny $X = \{x_1, \ldots, x_n\}$, množiny $Y = \{y_1, \ldots, x_m\}$ a relace definované dle: $\langle x_i, y_j\rangle \in I$, tehdy a pouze tehdy, obsahuje-li $\times$ na řádku i a sloupci j.

Definice 2 (Šipkové operátory)
Pro formální kontext $\langle X, Y, I\rangle$ jsou definovány operátory $^{\uparrow} : 2^X \to 2^Y$ a $^{\downarrow} : 2^Y \to 2^X$ tak, že pro každé $A \subseteq X$ a $B \subseteq Y$:

$$A^{\uparrow} = \{y \in Y \mid \text{pro každý } x \in A : \langle x, y\rangle \in I\}$$
$$B^{\downarrow} = \{x \in X \mid \text{pro každý } y \in B : \langle x, y\rangle \in I\}$$

Operátor $^{\uparrow}$ zjišťuje všechny společné atributy pro objekty z A. Naopak operátor $^{\downarrow}$ zjišťuje všechny objekty sdílející všechny atributy z B.

Příklad 3
Pro tabulku

	y_1	y_2	y_3	y_4
x_1		$\times$		$\times$
x_2			$\times$	$\times$
x_3	$\times$	$\times$		
x_4	$\times$		$\times$	

platí

$$\{x_1, x_2\}^{\uparrow} = \{y_4\}, \{x_3, x_4\}^{\uparrow} = \{y_1\},$$
$$\{x_3\}^{\uparrow} = \{y_1, y_2\},$$
$$X^{\uparrow} = \emptyset, \emptyset^{\uparrow} = Y,$$
$$\{y_3\}^{\downarrow} = \{x_2, x_4\}, \{y_3, y_4\}^{\downarrow} = \{x_2\},$$
$$\{y_1, y_3\}^{\downarrow} = \{x_4\},$$
$$\emptyset^{\downarrow} = X, Y^{\downarrow} = \emptyset.$$

Definice 4 (Formální koncept)
Formální koncept v $\langle X, Y, I\rangle$ je dvojice $\langle A, B\rangle$, kde $A \subseteq X$, $B \subseteq Y$ a platí:

$$A^{\uparrow} = B \wedge B^{\downarrow} = A$$

Množina A se nazývá *extent*, B se nazývá *intent*. Koncept je tvořen objekty z množiny A a atributy z množiny B. Atributy z B jsou právě všechny společné objektům z A a zároveň A jsou právě všechny objekty, co sdílí atributy z B.

Příklad 5 (Formální koncept)
Uvedená definice formalizuje naivní definici z předchozí podkapitoly a tedy $\langle${Rákos, Žába}, {a, b, c}$\rangle$ je konceptem, splňuje:

$$\{\text{Rákos, Žába}\}^{\uparrow} = \{a, b, c\} \wedge \{a, b, c\}^{\downarrow} = \{\text{Rákos, Žába}\}$$

Definice 6 (Uspořádání konceptů)
O konceptech $\langle A_1, B_1\rangle$ a $\langle A_2, B_2\rangle$ v kontextu $\langle X, Y, I\rangle$ řekneme, že

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle \textbf{ právě když } A_1 \subseteq A_2 (B_2 \subseteq B_1) \tag{1}$$

Relaci $\leq$ lze slovně interpretovat jako „být konkrétnější“. Pokud platí (1), pak koncept $\langle A_1, B_1\rangle$ je konkrétnější než koncept $\langle A_2, B_2\rangle$

Definice 7 (Svaz)
Nechť $\langle X, \leq\rangle$ je uspořádaná množina. Pokud pro každé $x, y \in X$ existuje $\sup(x, y)$ a $\inf(x, y)$, pak $\langle X, \leq\rangle$ se nazývá svaz. Pokud lze nalézt infima a suprema pro jakoukoliv podmnožinu množiny X, pak se jedná o úplný svaz.

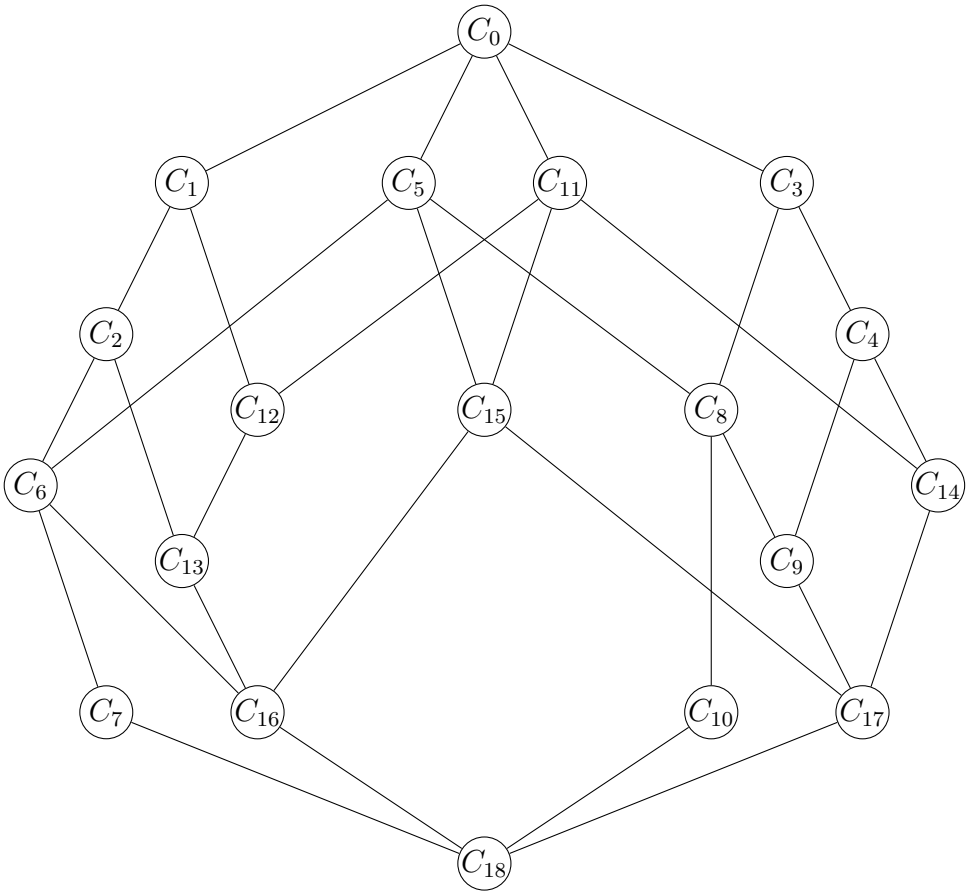
Definice 8 (Konceptuální svaz)
Pro kontext $\langle X, Y, I\rangle$ je definovaná množina všech formálních konceptů:

$$\mathcal{B}(X, Y, I) = \{\langle A, B\rangle \in 2^X \times 2^Y \mid A^{\uparrow} = B \wedge B^{\downarrow} = A\}.$$

Dvojice $\langle \mathcal{B}(X, Y, I), \leq\rangle$ se nazývá konceptuální svaz.

5.2.1 Vizualizace konceptuálního svazu

Zde jsou vypsány všechny formální koncepty z formálního kontextu prezentovaného v tabulce 1. Vizualizovány na obrázku 7.
$C_0 = \langle\{1, 2, 3, 4, 5, 6, 7, 8\}, \{a\}\rangle$, $C_1 = \langle\{2, 5, 6, 8\}, \{a, g\}\rangle$,
$C_2 = \langle\{2, 5, 8\}, \{a, g, h\}\rangle$, $C_3 = \langle\{1, 3, 4, 7\}, \{a, d\}\rangle$,
$C_4 = \langle\{1, 4, 7\}, \{a, d, f\}\rangle$, $C_5 = \langle\{3, 4, 5, 7, 8\}, \{a, c\}\rangle$,
$C_6 = \langle\{5, 8\}, \{a, c, g, h\}\rangle$, $C_7 = \langle\{5\}, \{a, c, g, h, i\}\rangle$,
$C_8 = \langle\{3, 4, 7\}, \{a, c, d\}\rangle$, $C_9 = \langle\{4, 7\}, \{a, c, d, f\}\rangle$,
$C_{10} = \langle\{3\}, \{a, c, d, e\}\rangle$, $C_{11} = \langle\{1, 2, 6, 7, 8\}, \{a, b\}\rangle$,
$C_{12} = \langle\{2, 6, 8\}, \{a, b, g\}\rangle$, $C_{13} = \langle\{2, 8\}, \{a, b, g, h\}\rangle$,



Obrázek 7: Konceptuální svaz zobrazený pomocí Hasseova diagramu

$C_{14} = \langle\{1,7\},\{a,b,d,f\}\rangle$, $C_{15} = \langle\{7,8\},\{a,b,c\}\rangle$,
$C_{16} = \langle\{8\},\{a,b,c,g,h\}\rangle$, $C_{17} = \langle\{7\},\{a,b,c,d,f\}\rangle$,
$C_{18} = \langle\{\},\{a,b,c,d,e,f,g,h,i\}\rangle$.

Věty, definice a další, obsažené v této kapitole citováno z [6], [7].

6 Faktorové koncepty

Nechť I je binarní matice s rozměry $n \times m$. Cílem je tuto matici rozložit na produkt matic $I = A \circ B$ binárních matic A a B o rozměrech $n \times k$ a $k \times m$, kde k je co nejmenší. Produkt binárních matic je definován následovně

$$(A \circ B)_{ij} = \bigvee_{l=1}^{k} A_{il} \cdot B_{lj},$$

kde $\bigvee$ označuje pravdivostní funkci logické disjunkce a $\cdot$ pravdivostní funkci logické konjunkce.

Příklad 9

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \circ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

V původní matici $I_{ij} = 1$ značí, že objekt i má atribut j, $A_{il} = 1$ znamená, že faktor l platí pro objekt i, nakonec dle $B_{li} = 1$ atribut j je jeden z projevů faktoru l. Dekompozice matice I do $A \circ B$ odpovídá objevu k faktorů, které definují data reprezentované maticí I.

6.1 Formální koncepty jako optimální faktory

Nechť I je binární matice. Cílem je tuto matici rozložit na produkt matic $I = A \circ B$ binárních matic A a B o rozměrech $n \times k$ a $k \times m$. Rozklad matice I na produkt

$$I = A_{\mathcal{F}} \circ B_{\mathcal{F}},$$

binárních matic $A_{\mathcal{F}}$ a $B_{\mathcal{F}}$ vytvořených z množiny formálních konceptů $\mathcal{F}$ spjatých s I. S I je spojený konceptuální svaz $\mathcal{B}(X, Y, I)$ s $X = \{1, \ldots, n\}$ a $Y = \{1, \ldots, m\}$. Nechť

$$\mathcal{F} = \{\langle A_1, B_1 \rangle, \ldots, \langle A_k, B_k \rangle\} \subseteq \mathcal{B}(X, Y, I),$$

$\mathcal{F}$ je tedy množina formálních konceptů z $\mathcal{B}(X, Y, I)$. Matice $A_{\mathcal{F}}$ a $B_{\mathcal{F}}$ o rozměry $n \times k$ a $k \times m$ jsou definovány následovně:

$$(A_{\mathcal{F}})_{il} = \begin{cases} 1 & \text{if } i \in A_l, \\ 0 & \text{if } i \notin A_l, \end{cases} \qquad \text{a} \qquad (B_{\mathcal{F}})_{lj} = \begin{cases} 1 & \text{if } j \in B_l, \\ 0 & \text{if } j \notin B_l, \end{cases}$$

pro $l = 1, \ldots, k$. Kde l-tý sloupec $(A_{\mathcal{F}})_{\_l}$ matice $A_{\mathcal{F}}$ sestává z charakteristického vektoru A_l a l-tý řádek $(B_{\mathcal{F}})_{l\_}$ matice $B_{\mathcal{F}}$ se skládá z charakteristického vektoru B_l.

Příklad 10

Pro danou matici I

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix},$$

jsou $\langle\{1,2,3\},\{1,2\}\rangle$ a $\langle\{1,2,3,4\},\{1\}\rangle$ formální koncepty z přidruženého konceptuálního svazu. Přidáním do

$$\mathcal{F} = \{\langle\{1,2,3\},\{1,2\}\rangle, \langle\{1,2,3,4\},\{1\}\rangle\},$$

vzniknou matice

$$(A_{\mathcal{F}}) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{pmatrix} \qquad \text{a} \qquad (B_{\mathcal{F}}) = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Zde ale $I \neq A_{\mathcal{F}} \circ B_{\mathcal{F}}$. Otázkou je zda pro každou I existuje nějaká množina $\mathcal{F} \in \mathcal{B}(X,Y,I)$ tak, že $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$. Odpovědí je následující věta.

Věta 11 (Univerzalita formálních konceptů jako faktorů)
Pro každé I existuji $\mathcal{F} \subseteq \mathcal{B}(X,Y,I)$ takové, že $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$.

Věta 12 (Optimalita formálních konceptů jako faktorů)
Nechť $I = A \circ B$ pro $n \times k$ a $k \times m$ binární matice A a B. Potom zde existuje množina $\mathcal{F} \subseteq \mathcal{B}(X,Y,I)$ formálních konceptů z I s vlastností

$$|\mathcal{F}| \leq k,$$

takové, že pro $n \times |\mathcal{F}|$ a $|\mathcal{F}| \times m$ binární matice $A_{\mathcal{F}}$ a $B_{\mathcal{F}}$ máme

$$I = A_{\mathcal{F}} \circ B_{\mathcal{F}}.$$

Rozklad pomocí formálních konceptů jako faktorů je optimální ve smyslu, že přináší co nejmenší počet faktorů.

Definice 13
Množina $\mathcal{F} \subseteq \mathcal{B}(X,Y,I)$ se nazývá množinou faktorových konceptů pokud $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$.

6.1.1 Mandatorní faktory

Předchozí věta ukazuje, že určité formální koncepty jsou mandatorní, musí být tedy obsaženy v každé množině $\mathcal{F}$ pro kterou $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$

$$\mathcal{O}(X,Y,I) = \{\langle\{x\}^{\uparrow\downarrow}, \{x\}^{\uparrow}\rangle \mid x \in X\},$$
$$\mathcal{A}(X,Y,I) = \{\langle\{y\}^{\uparrow\downarrow}, \{y\}^{\uparrow}\rangle \mid y \in Y\}.$$

Věta 14 (Mandatorní faktory)
Pokud $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$ *pro nějakou množinu* $\mathcal{F} \subseteq \mathcal{B}(X, Y, I)$ *pak* $\mathcal{O}(X, Y, I) \cap \mathcal{A}(X, Y, I) \subseteq \mathcal{F}$

Formální koncepty, které patří mezi objektové i atributové koncepty jsou mandatorní.

INPUT: I (Booleanovská matice)
OUTPUT: $\mathcal{F}$
$\mathcal{U} \leftarrow \{\langle i, j\rangle \mid I_{ij} = 1\}$
$\mathcal{F} \leftarrow \emptyset$
while $\mathcal{U} \neq \emptyset$ **do**
  $D \leftarrow \emptyset$
  $V \leftarrow 0$
  while existuje $j \notin D$ pro které $|D \oplus j| > V$ **do**
    vyber $j \notin D$ maximalizující $D \oplus j$
      $D \leftarrow (D \cup \{j\})^{\uparrow\downarrow}$
      $V \leftarrow 0$
  end while
  $C \leftarrow D^{\downarrow}$
  přidej $\langle C, D\rangle$ do $\mathcal{F}$
end while
for each $\langle i, j\rangle \in C \times D$ **do**
  odeber $\langle i, j\rangle$ z $\mathcal{U}$
end for

Algorithm 1: Algoritmus GreConD [8]

Konstrukce faktorových konceptů v tomto algoritmu spočívá v sekvenčním přidávání „slibných“ sloupců. Tato idea je založena na faktu, že každý formální koncept $\langle C, D\rangle$ může být vyjádřen jako $D = \bigcup_{y \in D} \{y\}^{\downarrow\uparrow}$. Dále je využito pozorování, že pokud $y \notin D$ pak $\langle (D \cup \{y\})^{\downarrow}, (D \cup \{y\})^{\downarrow\uparrow}\rangle$ je formální koncept s $D \subset D \cup \{y\})^{\downarrow\uparrow}$. Proto lze vytvořit libovolný formální koncept sekvenčním přidáváním $\{y\}^{\downarrow\uparrow}$ do prázdné množiny atributů. Zde se využívá žravého přístupu a vybírá se $y \in Y$ maximalizující

$$D \oplus y = ((D \cup \{y\})^{\downarrow} \times (D \cup \{y\})^{\downarrow\uparrow}) \cap \mathcal{U}.$$

Příklad 15

V prvním kroku algoritmu se vybere $1 \in Y$, tedy první sloupec, protože $|\emptyset \oplus 1| = 6$, což odpovídá maximální možné hodnotě v aktuálním kroku. Jelikož $\{1\}^{\downarrow} = \{1, 3\}$ a $\{1\}^{\downarrow\updownarrow} = \{1, 5, 6\}$, prvním vybraným faktorovým konceptem

$$\begin{pmatrix} \mathbf{1} & 0 & 1 & 0 & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \mathbf{1} & 1 & 0 & 1 & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Obrázek 8: Vstupní matice, zvýrazněn faktor koncept z prvního kroku

je $\langle\{1,3\},\{1,5,6\}\rangle$, zvýrazněn na obrázku 8. Žádný další atribut již nelze přidat k $\{1,5,6\}$, protože přidání by nevedlo ke zvýšení počtu jedniček v matici, pokrytých tímto faktorem. Následně je tento faktor z matice odebrán.

V dalším kroku je vybrán faktorový koncept $\langle\{2\}^{\downarrow},\{2\}^{\downarrow\updownarrow}\rangle = \langle\{3,5\},\{2,4,6\}\rangle$. Algoritmus dále pokračuje a v následujících krocích jsou voleny atributy na sloupcích 3 a 6. Po nesplnění $\mathcal{U} \neq \emptyset$ algoritmus skončí a je získána množina

$$\mathcal{F} = \{\langle\{1,3\},\{1,5,6\}\rangle, \langle\{3,5\},\{2,4,6\}\rangle, \langle\{1,2,4,5\},\{3\}\rangle\, \langle\{1,3,4,6\},\{6\}\rangle\},$$

indukující faktorizaci $A_{\mathcal{F}} \circ B_{\mathcal{F}}$:

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \circ \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

6.2 Implementace algoritmu GreConD

Algoritmus 1 byl jako všechny algoritmy, které jsou součástí této práce implementován v jazyce Java. Obsažen ve třídě FCF včetně pomocných metod. Při spuštění algoritmu se vstupní Booleovskou maticí, která je vytvořená z datasetů savců a globálních klimatických proměnných, je nalezeno na 307 faktorových konceptů. Vstupní matice je popsána viz obrázek 6.

Faktorizací došlo k získání matice $A_{\mathcal{F}}$ s rozměry 245×307 a matice $B_{\mathcal{F}}$ s rozměry $307 \times 395\,760$, reprezentací dat ze vstupní matice pomocí uvedených matic by k žádné úspoře prostoru nedošlo. Nicméně hlavním cíle této práce bylo tyto nalezené faktorové koncepty, potenciálně zajímavé shluky dat podrobit analýze.

Jak algoritmus běží, nachází jednotlivé faktorové koncepty, které přidává do seznamu, konec nastává při $\mathcal{U} \neq \emptyset$. Universum je zde implementováno pomocí kolekce *HashMap*, množinu klíčů tvoří indexy sloupců, k indexu sloupce jsou namapovány indexy řádků. K indexu j sloupce je namapován index i řádku, pokud $I_{ij} = 1$. Ověření, zda universum není prázdné, provedené před každou iterací je řešeno pomocí `universumTable.isEmpty()`.

Faktory jsou přidávány do seznamu seznamů, který je implementován prostřednictvím `List<List<Integer» formalFactors`, výsledná data mají tvar $[[C_1], [D_1], \ldots [C_n], [D_n]]$.

V uvedeném formátu jsou ukládána na disk, do souboru `factorsRaw.txt`. Pro další analýzu, dojde k vytvoření dalších třech souborů, které z těchto dat vycházejí.

- `factorsInDetails.txt` - čitelnější verze souboru `factorsRaw.txt`, místo indexů řádků a sloupců jsou dosazeny hodnoty, na které indexy ukazují. Faktor je zde očíslován, následně je vypsán extent ve formě latinských názvů savců, dále intent, což jsou spjaté oblasti s jejich hodnotami klimatických proměnných. Ukázka viz zdrojový kód 10.

- `averages.txt` - postupně podle faktorů jsou v tomto souboru vypsány průměrné měsíční hodnoty klimatických proměnných pro jednotlivé savce z extentu, hodnoty jsou tedy průměrovány ze všech oblastí, kde se daný savec vyskytuje a přes jednotlivé měsíce. Vždy na konci řádku pro jednotlivou klimatickou proměnnou je vypsána i její roční průměrná hodnota. Nakonec jsou pro porovnání vypsány průměrné měsíční a roční hodnoty proměnných ze všech oblastí faktoru.

- `differences.txt` - v tomto souboru jsou vypsány rozdíly mezi průměrnými hodnotami proměných v měsících pro jednotlivé savce z extentu faktoru a hodnotami proměných zprůměrovaných ze všech oblastí daného faktoru.

```
1 ------------------FACTOR NUMBER:306 ------------------
2 <[Erinaceus+concolor, Crocidura+suaveolens, Suncus+etruscus,
    Rhinolophus+hipposideros, Lepus+europaeus, Meriones+tristrami,
    Apodemus+mystacinus, Apodemus+sylvaticus, Rattus+norvegicus,
    Rattus+rattus, Mus+domesticus, Vulpes+vulpes, Martes+foina,
    Monachus+monachus]
3 [35SNA1, tmin: 6.39, 6.29, 8.07, 11.16, 15.34, 19.76, 23.63, 23.55,
    20.93, 16.79, 12.5, 9.42, tmax: 15.83, 15.73, 17.51, 20.6,
    24.78, 29.2, 30.02, 29.95, 27.33, 23.18, 18.9, 15.81, tavg:
    11.11, 11.01, 12.79, 15.88, 20.06, 24.48, 26.83, 26.75, 24.13,
    19.98, 15.7, 12.62, prec: 159.83, 121.53, 80.96, 37.87, 18.01,
    7.37, 1.43, 3.71, 13.45, 52.92, 87.11, 154.52, srad: 8983.5,
    11977.2, 16237.73, 20771.76, 24559.7, 28245.97, 28627.63,
    26302.41, 21726.16, 15421.56, 10920.39, 8428.72, wind: 4.07,
    4.27, 3.99, 3.72, 3.56, 4.08, 4.8, 4.63, 3.96, 3.62, 3.53, 3.9,
    vapr: 0.89, 0.88, 0.98, 1.15, 1.42, 1.65, 1.82, 1.92, 1.68,
    1.46, 1.21, 1.02,
4 ]>
```

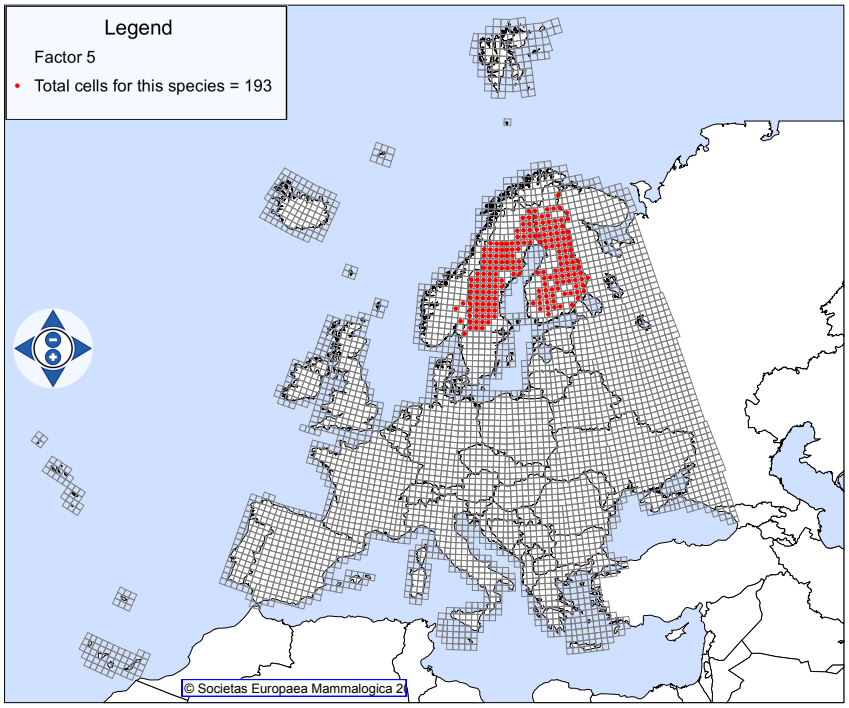
Zdrojový kód 10: Ukázka ze souboru factorsDetail

Věty, definice a další, obsažené v této kapitole citováno z [8].

6.3 Vizualizace faktorů

Na základě údajů z výstupních textových souborů je velice obtížné představit si zasazení výskytu savců do kontextu našeho světa. Umístění je zde popsáno pouze identifikátory jednotlivých oblastí, které vedou na pole souřadnic, reprezentující tuto oblast. Proto je součástí práce automatická tvorba obrázků, kde jsou tyto oblasti faktoru zvýrazněny na mapě Evropy. Jedná se doplněk k výstupním textovým souborům, které poskytují hodnoty klimatických proměnných u daných oblastí.

Jako výchozí šablona byl použit obrázek z webových stránek evropské databáze savců. Obrázek vizualizuje výskyt savce, viz obrázek 1. Šablona tedy už obsahuje vykreslenou mapu Evropy, kterou překrývá CGRS mřížka a dále obsahuje také legendu. Červená kolečka v mřížce představují oblasti obsažené v daném faktoru. V legendě je vypsáno o jaký faktor se jedná a doplňující informace o celkovém počtu obsažených oblastí.



Obrázek 9: Oblasti faktoru

Zdrojový kód stojící za vizualizací pokrývá třída *SvgCreator* a výsledné obrázky jsou uloženy ve složce `img/factors/`. Konstruktor třídy **SvgCreator** vyžaduje identifikátory oblastí určené k vykreslení, číslo faktoru a cestu, kde výsledný obrázek uložit. SVG obrázek definuje grafiku v XML formátu, tudíž se k šabloně obrázku přistupuje stejně jako ke standardnímu souboru XML. Metodě **addAreas** je předán seznam identifikátorů oblastí k vykreslení. Díky identifikátoru jsou snadno vyhledány „cesty“, které popisují zobrazení jednotlivé buňky mřížky, viz zdrojový kód 11.

Pro každou buňku je vypočítán střed, pro správné umístění červeného kolečka. Definice kolečka viz zdrojový kód 12. Všechna kolečka na mapě vychází ze vzoru, který je v obrázku skrytý, před umístěním do buňky je vzorové kolečko naklonováno a vzniklému klonu se nastaví správné umístění pomocí atributů x a y.

```
1 <path d="M55803 25834l-142-485l192-56l186-73l184 470l-207 81l-213 63
     z" id="37VFF3"/>
```

Zdrojový kód 11: Reprezentace oblasti

```
1 <circle fill="red" id="post70_1" r="200"/>
```

Zdrojový kód 12: Kolečko vzor

Před uložením na disk dojde k vyplnění legendy, obstarává metoda *setLegend*. Vyhledá element reprezentující legendu, která obsahuje tři potomky `<text>`, první obsahuje nadpis legendy, do druhého se vloží číslo faktoru a do třetího se vloží informace o počtu zobrazených koleček na mapě.

7 Analýza faktorů

Ve vstupní matici 6 bylo pomocí algoritmu 1 nalezeno celkem na 307 faktorů. Detailně bylo analyzováno prvních 25 faktorů, každý faktor se nicméně interpretovat nepodařilo. Nejzajímavější faktory z analyzovaných jsou rozebrány na následujících stranách.

Faktory byly zkoumány na základě vzájemných vztahů jednotlivých obsažených savců, ať už z hlediska vztahů potravních, konkurenčních, symbiotických a dalších. Dále byl brán v potaz výskyt každého savce vůči oblastem faktoru. Zahrnuty byly také hodnoty klimatických proměnných vázané na oblasti výskytu savců tak na oblasti faktoru.

Pro každou oblast existuje záznam o minimální, maximální, průměrné teplotě a také záznam o srážkách, intenzitě slunečního záření, rychlosti větru a tlaku vodní páry. Hodnoty byly získávány měřením ve sledovovaném období 1970-2000. Při analýze byly tyto hodnoty u oblastí výskytu savců porovnávány vůči hodnotám náležících oblastem faktoru. Nejen v kontextu celého roku, ale i v kontextu jednotlivých měsíců. Ve stejném duchu byly tyto hodnoty porovnávány i pro jednotlivé savce mezi sebou.

Na nalezené oblasti se nahlíželo i z pohledu geologického umístění. Přesněji z pohledu příslušnosti oblastí k podnebím panujících na evropském kontinentu.

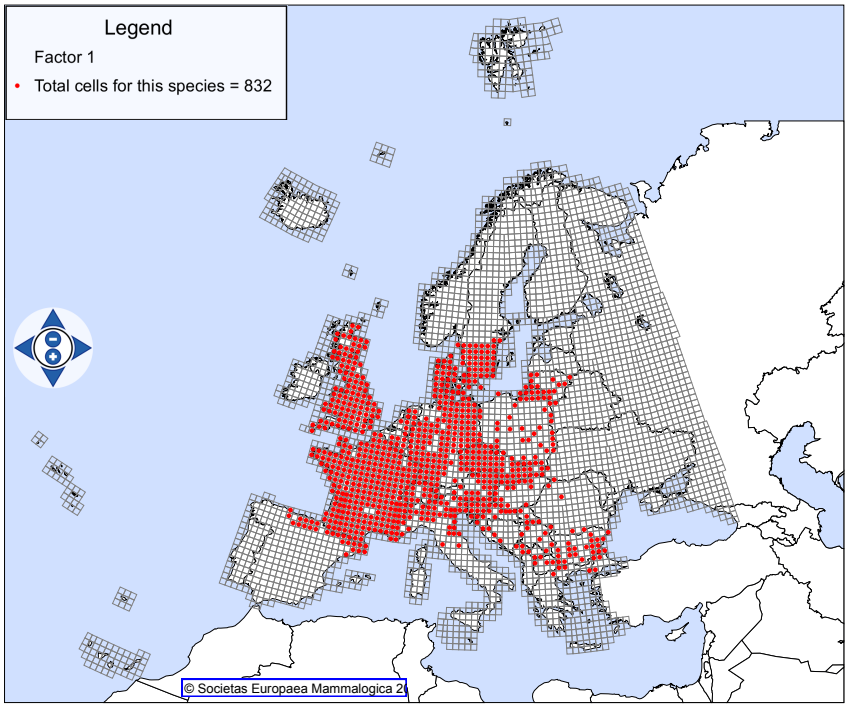
7.1 Faktor 1

Ve faktoru 1 jsou zahrnuti tito savci: krtek obecný (*Talpa europaea*), zajíc polní (*Lepus europaeus*), myšice křovinná (*Apodemus sylvaticus*), liška obecná (*Vulpes vulpes*), lasice kolčava (*Mustela nivalis*) a jezevec lesní (*Meles meles*).

Oblasti pro tento faktor se drží v pásmu mírného a kontitentálního podnebí dle Köppenovi klasifikace podnebí [9]. Severně jsou oblasti ohraničeny subpolárním podnebím na jihu Švédska, z jihu pak ohraničeny na severu Španělska a Itálie podnebím semidiárním.

Větší podobnosti mezi oblastmi výskytu jednotlivých savců a oblastmi faktoru lze hledat pouze stěží. Jsou tu savci mající k podmínkám panujícím v oblastech faktoru velice blízko. Například průměrná měsíční v oblastech výskytu zajíce polního je nižší pouze o 0,03°. U lišky obecné pak o −0,6° nižší a naopak u myšice křovinné, preferující teplejší oblasti, je průměrná měsíční teplota o 0,65° vyšší. U dalších proměnných jsou rozdíly obdobné.

Mezi uvedenými savci existují potravní vztahy. Před liškou obecnou není v bezpečí zajíc polní [10], myšice křovinná [11], dále také lasice kolčava [12]. Myšice křovinná je také potravou pro lasici kolčavu, lasice je schopna ulovit krtka obecného a využít jeho vyhloubené doupě [13].



Obrázek 10: Oblasti faktoru 1

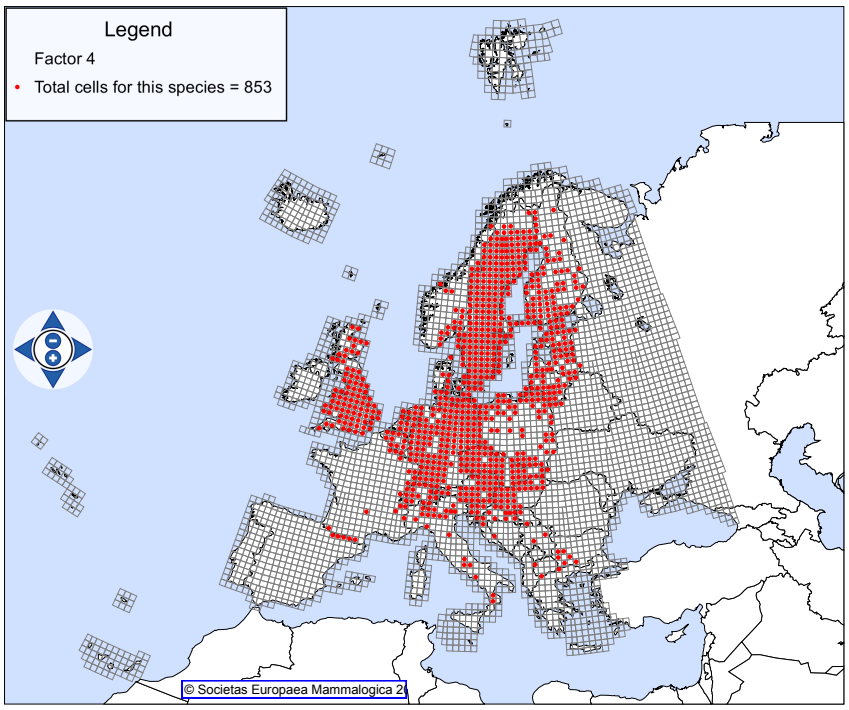
7.2 Faktor 4

Zajímavý je faktor s číslem 4. Obsahuje následující savce: rejsek obecný (*Sorex araneus*), rejsek malý (*Sorex minutus*), norník rudý (*Clethrionomys glareolus*), hryzec vodní (*Arvicola terrestris*), liška obecná (*Vulpes vulpes*). Oblasti spadající pod faktor leží v mírném podnebném pásmu, subpolárním podnebném pásmu a částečně zasahují i do subtropického pásma v Itálii. Průměrné klimatické hodnoty naměřené v oblastech výskytu u jednotlivých savců se značně liší od hodnot oblasti faktoru. Výskyt rejska obecného se nejvíce překrývá s oblastí faktoru, tudíž zde se hodnoty klimatických proměnných od průměru oblasti liší nejméně. V klimatických podmínkách by se podobnost hledala pouze stěží.

Z hlediska potravních vztahů lze savce rozdělit na

- hlodavci - rejsek obecný, rejsek malý, norník rudý, hryzec vodní.
- šelmy - liška obecná.

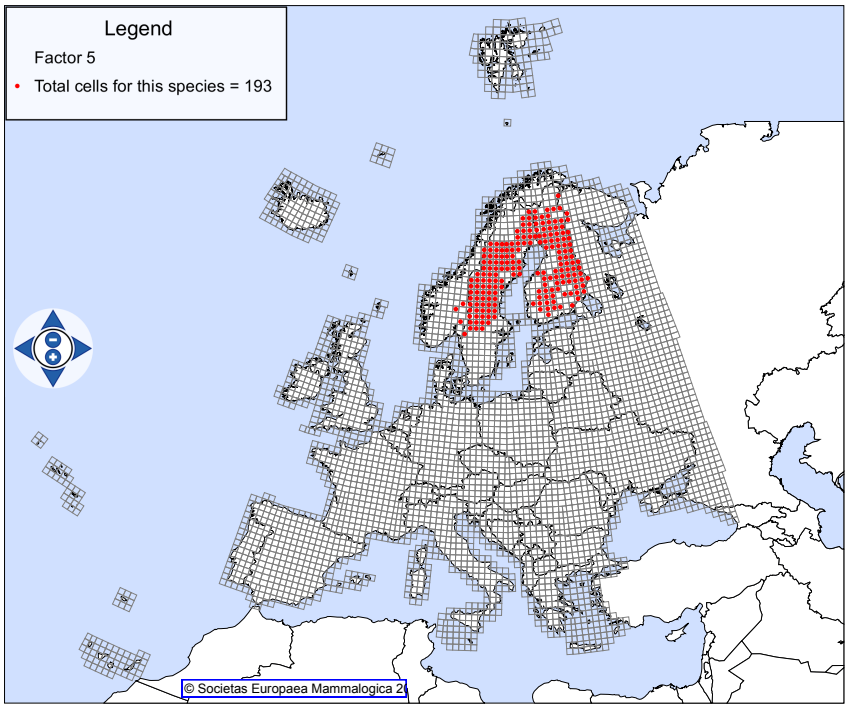
Přičemž všichni uvedení hlodavci patří mezi potravní zdroje lišky obecné [14][15][16][17]. Uvedený faktor lze tedy nazvat „Potravní zdroje lišky obecné mezi hlodavci“.



Obrázek 11: Oblasti faktoru 4

7.3 Faktor 5

Ve faktoru 5 je obsaženo 10 savců a 193 oblastí. Mezi dotyčné savce patří zajíc bělák (*Lepus timidus*), veverka obecná (*Sciurus vulgaris*), lumík lesní (*Myopus schisticolor*), liška obecná(*Vulpes vulpes*), medvěd hnědý (*Ursus arctos*), lasice hranostaj (*Mustela erminea*), lasice kolčava (*Mustela nivalis*), norek americký (*Mustela vison*), kuna lesní (*Martes martes*) a los evropský (*Alces alces*).



Obrázek 12: Oblasti faktoru 5

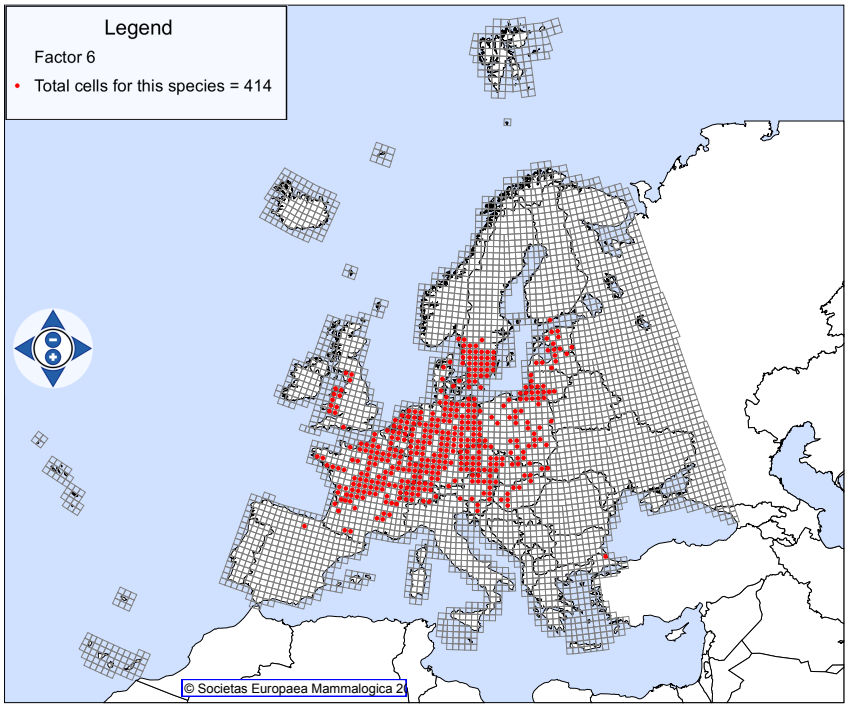
Pro tento faktor lze za nejvíce přiléhající název zvolit „Savci subpolárního pásma“, či ještě konkrétnější „Skandinávští savci“. Nicméně nazvat faktor například „Savci výhradně obývající Skandinávii“ nelze. Jediný savec výhradně obývající Skandinávii nebo subpolární pásmo je lumík lesní, hodnoty klimatických proměnných z oblastí výskytu lumíka se příliš neliší od hodnot pro oblast faktoru. Dále hodnoty u medvěda hnědého se liší více např. průměrná teplota v měsíci lednu až o 3,63°, protože mezi země výskytu medvěda hnědého patří i například Slovensko a Rumunsko. U silně invazivního druhu jakýmž je liška se liší průměrná teplota v únoru až o 10,28°.

Potravní vztahy mezi savci jsou také zajímavé. Na vrcholu potravního řetězce je liška obecná spolu s medvědem hnědým, všichni ostatní savci patří mezi kořist minimálně jednoho z předchozích predátorů. Veverka obecná se může navíc stát kořistí lasice hranostaj, lasice kolčavy, norka amerického a kuny lesní [18][19][12][20][21].

7.4 Faktor 6

Zde nejpočetnější skupinu tvoří netopýři: netopýr vodní (*Myotis daubentonii*), netopýr řasnatý (*Myotis nattereri*), netopýr rezavý (*Nyctalus noctula*), netopýr ušatý (*Plecotus auritus*). Dále je zde přítomen zajíc polní (*Lepus europaeus*), veverka obecná (*Sciurus vulgaris*) a tchoř tmavý (*Mustela putorius*).

Z hlediska klimatických hodnot oblasti faktoru pokrývají mírné a boreální (kontinentální) podnebí, konkrétně mírné oceánické podnebí (Cfb) a vlhké kontinentální podnebí (Dfb), dle Köppenovy klasifikace podnebí [9]. Oba tyto klimatické typy spojují teplá léta. Jsou zde zastoupeny i hraniční oblasti těchto typů, od jihu jsou to oblasti okolo španělsko–francouzské hranice. Od severu pak oblasti na jihu Švédska a Finska. Od průměrných teplot se nejvíce odchyluje veverka obecná, průměrné měsíční teploty panující na oblastech jejího výskytu se liší až o $-1{,}56°$, tchoř tmavý obecně obývá sušší oblasti, srážkovost v měsíci červenci je nižší o 11,51 mm.



Obrázek 13: Oblasti faktoru 6

Savce lze opět rozdělit na predátory a potenciální kořist. Tchoř tmavý je schopen ulovit zajíce polního [13]. Veverka obecná je lovena pouze vzácně [19]. Potravní vztahy mezi tchořem a netopýry nalezeny nebyly.

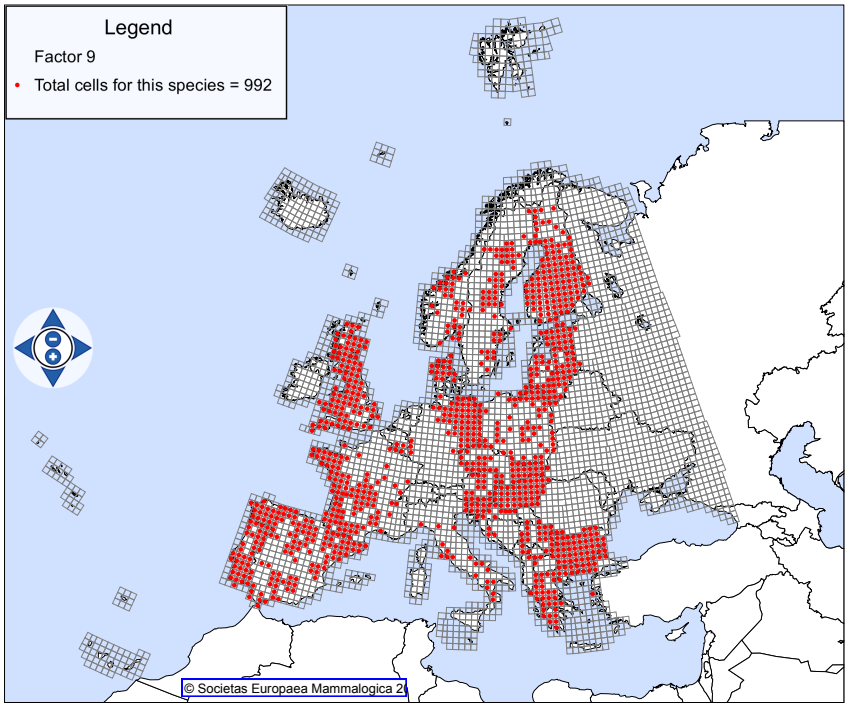
7.5 Faktor 9

Faktor 9 sestává pouze z šelem *(Carnivora)*. Z čeledi lasicovitých *(Mustelidae)* je zde zastoupen jezevec lesní *(Meles meles)*, lasice kolčava *Mustela nivalis* a vydra říční *(Lutra lutra)*. Z psovitých je přítomna liška obecná *(Vulpes vulpes)*.

Oblastí obsažených je zde na 992. Objevují se zde klimatická pásma od studeného stepního podnebí ve Španělsku (Bsk) až po subpolární podnebí (Dfc) ve Finsku.

Průměrné měsíční teploty na oblasti výskytu jednotlivých šelem se příliš neliší od průměrných teplot území faktoru, přes všechny měsíce pouze v rozmezí $-0{,}38°$ až $0{,}41°$. Srážkovost se průměrně liší od 0,22 mm až do 2,03 mm. Hodnoty průměrného tlaku vodní páry na územích obývaných jednotlivými savci jsou prakticky totožné s průměrnými hodnotami této veličiny u oblastí tohoto faktoru. Všechny tyto rozšířené šelmy žijí ve velmi podobných klimatických podmínkách.

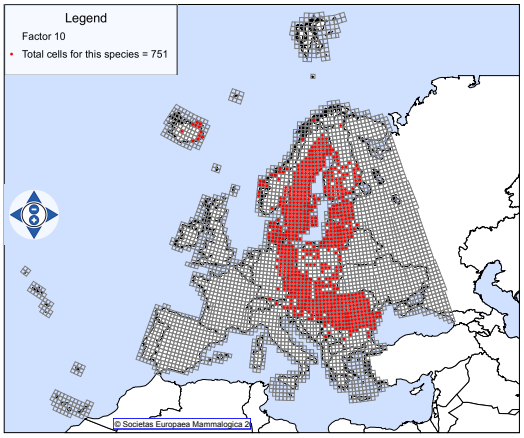
Potravní vztahy zde nejsou významné. Pouze lasice kolčava může být lovena liškou obecnou. Další potravní vztahy lze nalézt už jen stěží. Zajímavé je také možné sdílení nory mezi liškou obecnou a jezevcem lesním [22].



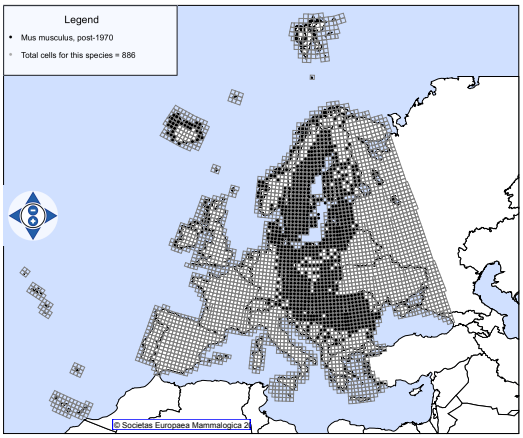
Obrázek 14: Oblasti faktoru 9

7.6 Faktor 10

Extent faktorového konceptu o dvou savcích: myš domácí (*Mus musculus*), potkan obecný (*Rattus norvegicus*). Intent tvoří 751 nalezených oblastí s jejich hodnotami. Oblasti se rozprostírají přes vlhké kontinentální podnebí s horkými léty (Dfa), vlhké kontinentální podnebí s teplými léty (Dfb) a subpolární podnebí s chladými léty a studenými zimami. Pro zmíněná podnebí platí, že teplota nejteplejšího měsíce je nad 10° a teplota nejstudenějšího měsíce pod −3°.



Obrázek 15: Oblasti faktoru 10



Obrázek 16: Výskyt myši domácí

Nalezené oblasti pokrývají více výskyt myši domácí než potkana obecného. Potkan má zaznamenaný výskyt na 1800 oblastech prakticky po celé Evropě. Tuto skutečnost potvrzují i data v souboru `averages.txt`. Například průměrná lednová teplota na oblastech výskytu potkana je o 4,02° vyšší než průměrná lednová teplota u oblastí faktoru. U myši je o 0,1° nižší. Dále srážkový úhrn pro leden je u potkana o 21,17 mm vyšší, u myši pouze 0,23 mm vyšší. Obdobně platí pro všechny sledované proměnné.

Faktor může nést název „Společný výskyt myši domácí a potkana obecného v oblastech s kontinentálním podnebím“. Zajímavé je také „vražedné chování“ potkana obecného vůči myši domácí. Až 70 % divoce žijících potkanů zabíjí myši. Vraždění není iniciováno pouze za účelem potlačení konkurence, potkan část myši zkonzumuje (mozek, hrudní tuk, vnitřnosti) [23].

7.7 Faktor 11

Faktor obsahující pouze tři savce: ježek západní (*Erinaceus europaeus*), veverka obecná (*Sciurus vulgaris*) a liška obecná (*Vulpes vulpes*). První savec faktoru je hmyzožravý, druhý hlodavec a třetí je šelma.

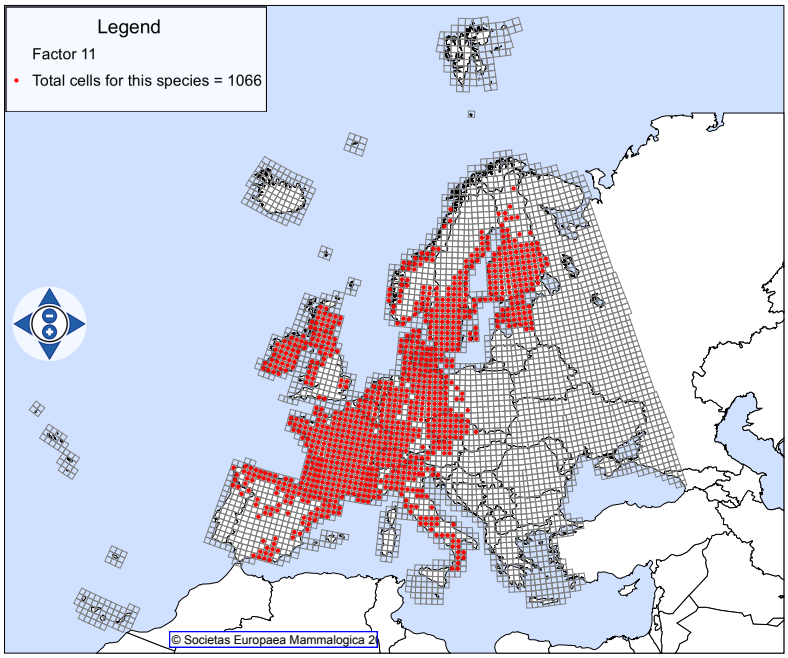
Obsažených je 1066 oblastí. Toto rozsáhlé území se rozprostírá od suchého podnebí přes mírné a kontinentální podnebí až k podnebí studenému. Jsou zde zastoupena všechna podnebí Evropy.

Ježek západní i veverka obecná se mohou stát kořistí lišky obecné. Ovšem nepatří mezi hlavní kořist lišky, ježek je často konzumován až ve formě mršiny[24], zanechané při kolizi ježka s automobilem.

Průměrné hodnoty klimatických proměnných v oblastech výskytu savců se od průměrné hodnoty na nalezených oblastech příliš neliší, přesto rozdíly existují. Nalezené oblasti nejvíc incidují s oblastmi výskytu ježka západního, ale není zde zastoupen jižnější výskyt ježka na Sicílii a v Portugalsku. Z toho pramení například o 1,23° teplejší leden u výskytu ježka a také vyšší solární radiace o 3000 kJ m^{-2} day^{-1}.

Z výskytu lišky zde naopak chybí severní výskyt v Norsku a výskyt na Balkánském poloostrově. Výskyt lišky oproti nalezeným oblastem provází lehce nižší průměrná teplota v lednu −0,4° a lehce vyšší v červenci o 0,62°. Srážky zde jsou nižší po všechny měsíce v roce, mezi −2,92 mm a −7,94 mm. Z výskytu veverky obecné chybí již výhradně severní oblasti, tudíž průměrná teplota oproti oblastem faktoru je zde nižší, srážky méně hojnější i solární radiace je nižší.

Faktor je zajímavý z pohledu podnebí, kdy jsou zde zastoupeny všechny možné, které mohou panovat na Evropském kontinentu.



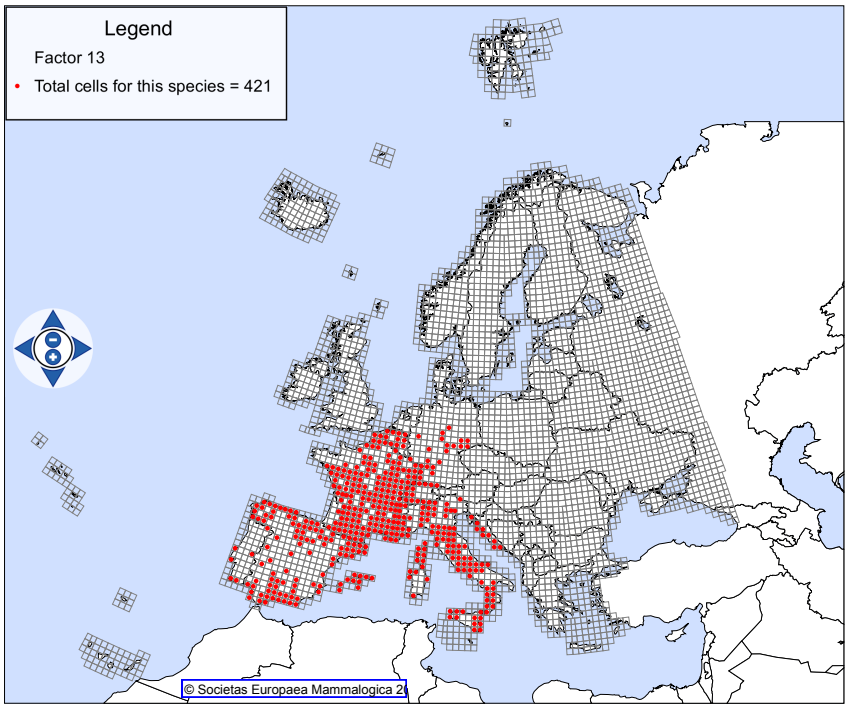
Obrázek 17: Oblasti faktoru 11

7.8 Faktor 13

Faktor 13 je pozoruhodný ze dvou důvodů. Zaprvé, všichni dotčení savci patří mezi hlodavce, myšice křovinná (*Apodemus sylvaticus*), krysa obecná (*Rattus rattus*), myš domácí (*Mus domesticus*) z čeledi myšovitých a plch zahradní (*Eliomys quercinus*) z čeledi plchovitých.

Zadruhé, oblasti pokrývají mírné podnebí - mírné oceánické podnebí (Cfb), vlhké subtropické podnebí (Cfa) a středomořské podnebí s teplými léty (Csb) a lze pozorovat „díru“ na území Španělska a Portugalska, v těchto faktorem neobsažených oblastech panují podmínky odpovídající suchému podnebí, klimatickému typu (Bsk).

Průměrně na oblastech faktoru naprší měsíčně 60,66 mm, srážky na výskytu jednotlivých savců se pohybují v rozmezí 57,75 mm - 61,82 mm. Ostatní klimatické hodnoty u oblastí výskytu jednotlivých savců se liší více. Myšice křovinná se vyskytuje na místech s vyšší průměrnou rychlostí větru až o 0,54 m/s^{-1} a zároveň nižší teplotou: −1,81° od měsíčního průměru 11,49° oblastí faktoru. Možný název pro tento faktor „Hlodavci mírného klimatu západní a jižní Evropy“ či „Hlodavci oblastí s průměrným měsíčním srážkovým úhrnem 57 mm - 62 mm“.



Obrázek 18: Oblasti faktoru 13

7.9 Faktor 14

Savci pro tento faktor: rejsek obecný (*Sorex araneus*), hraboš mokřadní (*Microtus agrestis*), lasice hranostaj (*Mustela erminea*), norek americký (*Mustela nivalis*) a liška obecná (*Vulpes vulpes*).

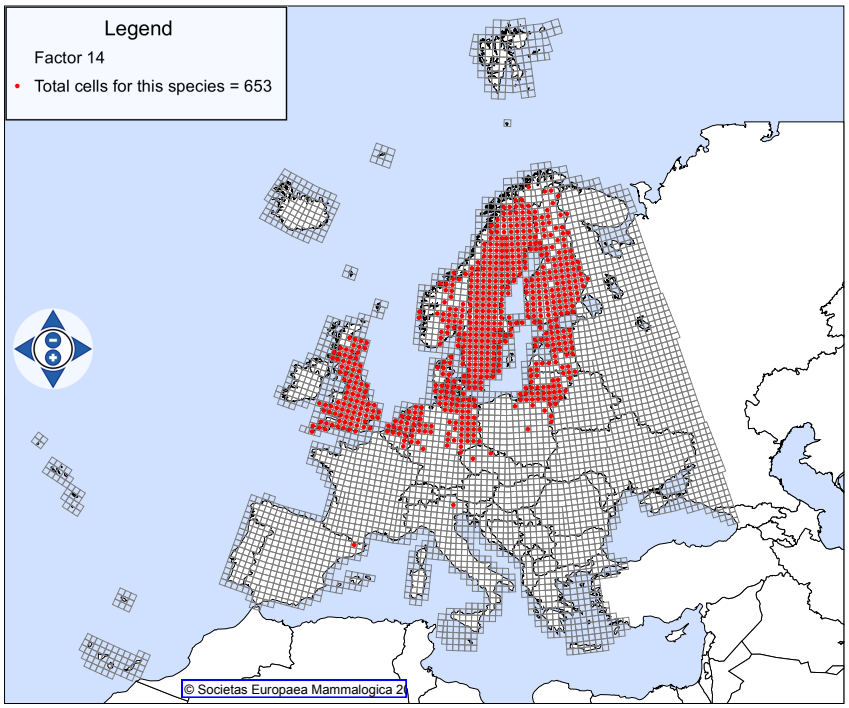
Celkem 653 oblastí, nejvíce zastoupeno kontinentální podnebí (mírné studené), konkrétní klimatické typy jsou vlhké kontinenální podnebí (Dfb) a subpolární podnebí (Dfc) v menší míře zastoupeno i mírné podnebí (mírné teplé) v podobě mírného oceánického podnebí (Cfb).

Potravní vztahy lze rozdělit na tři úrovně:

1. Úroveň, nachází se zde rejsek obecný a hraboš mokřadní, členové této úrovně jsou potravou pro savce ve vyšších úrovních [25][14][26].

2. Úroveň, dva zástupci lasicovitých, norek americký a lasice hranostaj, loví savce 1. úrovně ale jsou loveni savci vyšší úrovně [12].

3. Úroveň, jediný členem je liška obecná, v rámci tohoto faktoru není ničí potravou a všechny ostatní savce aktivně loví.



Obrázek 19: Oblasti faktoru 14

K průměrným klimatickým hodnotám oblasti faktoru má nejblíže norek americký, který se v podstatě vyskytuje pouze v oblastech, které odpovídají podnebím tohoto faktoru.

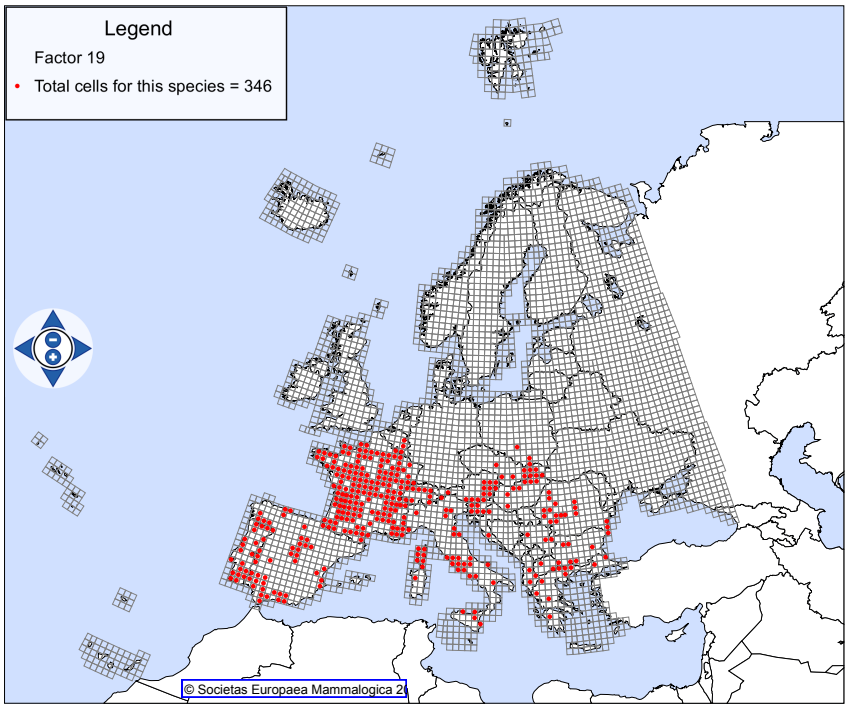
7.10 Faktor 19

Zde byli nalezeni pouze netopýři: vrápenec velký (*Rhinolophus ferrumequinum*), vrápenec malý (*Rhinolophus hipposideros*), netopýr velký (*Myotis myotis*). Zahrnuto na 346 oblastí s mírným podnebím.

Vrápenci žijí v oblastech s velmi podobnými klimatickými podmínkami. V nalezených oblastech je průměrná měsíční teplota 10,82°, průměrná měsíční teplota na oblastech výskytu vrápenců se liší pouze o 0,3°. Větrnost se pak liší o $0{,}12\,\mathrm{m/s^{-1}}$ od průměrné oblastní $2{,}86\,\mathrm{m/s^{-1}}$ a je shodná pro oba vrápence. Tlak vodní páry se odchyluje zanedbatelně.

Netopýr velký se vyskytuje i severněji od oblastí faktoru, například v Německu a Polsku. Průměrná měsíční teplota na oblastech jeho výskytu je tedy přirozeně chladnější, až o −0,9°. Ostatní proměnné se příliš neliší, obdobně jako u vrápenců. Některé mají dokonce k oblastním naopak blíže. Průměrná měsíční rychlost větru se liší o $0{,}02\,\mathrm{m/s^{-1}}$ od průměru nalezených oblastí.

Uvedenou skupinu lze pojmenovat „Létající savci“ nebo „Létající savci mírného podnebí“. Potravní vztahy mezi sebou netopýři nemají, navzájem se neloví, všichni se živí hmyzem.



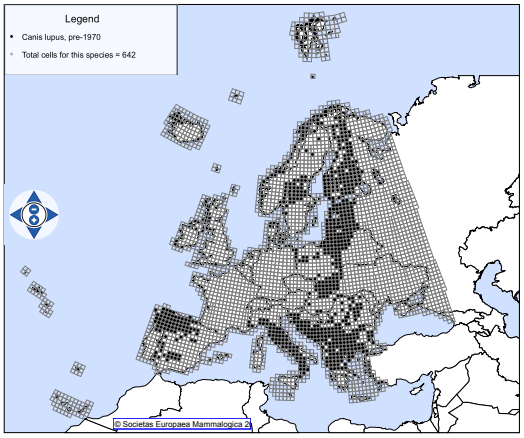
Obrázek 20: Oblasti faktoru 19

7.11 Faktor 20

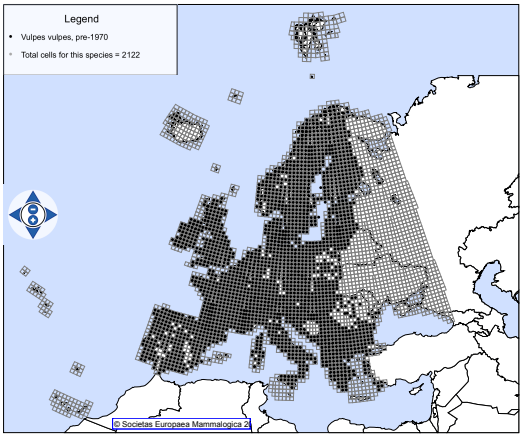
Faktor psovitých šelem: vlk obecný (*Canis lupus*) a liška obecná (*Vulpes vulpes*). Oblasti nalezeného faktoru víceméně odpovídají výskytu vlka obecného, klimatické hodnoty oblasti faktoru a hodnoty výskytu vlka jsou totožné. Hodnoty výskytu lišky se liší od hodnot oblastech faktoru výrazněji. Například průměrná lednová rychlost větru v oblastech faktoru je $3{,}02\,\mathrm{m/s^{-1}}$ u lišky pak $3{,}72\,\mathrm{m/s^{-1}}$. I dle ostatních klimatických proměnných lze dovodit, že tito savci se nacházejí v oblastech s rozdílnými podmínkami. Jak lze vidět z obrázku 22, liška obecná je velmi adaptabilní.

Z hlediska potravních vztahů může být liška kořistí vlka. Nicméně výskyt vlka může být pro lišku výhodný, obzvláště v zimních měsících, kdy vlčí smečka zanechá zbytky z větší kořisti, kterou by liška nebyla schopna ulovit [27].

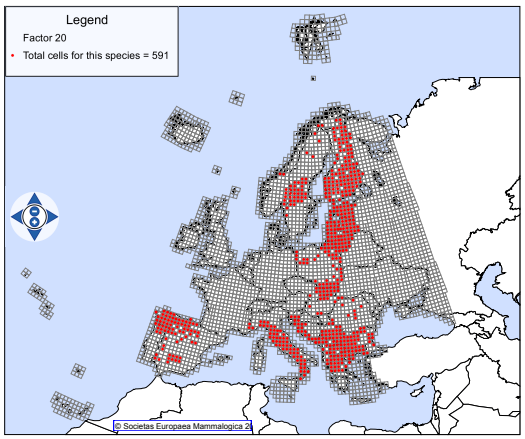
Název pro nalezený faktor zní: „Společné soužití vlka obecného a lišky obecné v Evropě“.



Obrázek 21: Výskyt vlka obecného



Obrázek 22: Výskyt lišky obecné



Obrázek 23: Oblasti faktoru 20

Závěr

Hlavním cílem diplomové práce byla extrakce dat z EMMA databáze (databáze evropské fauny) a dat z databáze WorldClim (klimatická databáze). Výsledkem této extrakce je dataset obsahující data z obou databází ve formě Booleovské matice. Tento cíl byl tedy naplněn a vzniklý dataset mohl být analyzován pomocí vybrané metody rozkladu matic.

Analýza výsledného datasetu byla dalším cílem této práce. Zvolená analytická metoda GreConD byla implementována v programovacím jazyce Java. Prostřednictvím této metody bylo v datasetu nalezeno 307 faktorových konceptů.

Nalezené faktorové koncepty byly dále zkoumány a vybrané se podařilo interpretovat. Mezi savci, kteří se vyskytují v jednotlivých faktorových konceptech byly zjištěny zajímavé potravní, konkurenční a další vztahy. Rovněž byly nalezeny a ukázany vztahy mezi savci a oblastmi výskytu, spolu s jejich klimatickými podmínkami. Tyto vztahy dále obsahují preference určitých specifických klimatických podmínek některými savci. Případně je pouze u jistých savců ukázána jejich adaptabilita vůči dostupným zdrojům potravy, či vůči klimatickým podmínkám. Závěrem byly nalezené faktorové koncepty vizualizovány na mapě Evropy.

V budoucnu je možné do výsledného datasetu zahrnout další informace. Například údaje o hustotě lidského osídlení v oblastech, údaje o střední nadmořské výšce a další. Nabízí se také možnost aplikovat jiné metody analýzy dat nebo zkoumat odlišnou geografickou oblast.

Conclusions

The aim of the study was the extraction of the data from the EMMA database (European fauna database) and from the WorldClim database (climatic database). The aim of this extraction was to create a dataset from both databases in the form of Boolean matrix. The aim was fulfilled and the created dataset was analysed using the matrix decomposition method.

The secondary aim of this study was the analysis of the result dataset. The chosen analytical method GreConD was implemented in the programming language Java. By using this method it was possible to find 307 factor concepts in the dataset.

The found factor concepts were further analysed and several chosen concepts were interpreted. The mammals that were found in the individual factor concepts showed an interesting competitive, food related and other relationships. The relationships between mammals and the area of occurrence were found and presented as well as the specific climatic conditions. These relationships include the specific climatic condition preferences of some of the mammal groups. The adaptability to different climatic conditions and available food sources was shown by certain mammal species. The found factor concepts were visualised on the map of Europe.

In the future it would be possible to include the altitude data, the density of human population in the area and other parameters. The analysis could also be carried out by using a different analytical method or by analysing a different geographical area.

A Popis SW

Zdrojový kód je rozdělen na dva balíčky `ExtractionTool` a `FactorFinder`. Stěžejní kód balíčku `ExtractionTool`:

- `MammalsExtractor` - přistupuje k webové stránce databáze evropských savců získá seznam všech dostupných savců z HTML struktury webové stránky. Následně předává serverové PHP metodě názvy jednotlivých savců. Tímto způsobem jsou získán výskyt savců ve formě identifikátorů oblasti. Při úspěchu uloží získaná data do souboru `mammals.xml`. Pokud je program opět spuštěn, seznam savců získá z uvedeného souboru.
- `MammalsMapExtractor` - identifikátorům oblastí přiřadí GPS koordináty. Informace o těchto koordinátech jsou k nalezení v souboru `CGRSJSON` na webových stránkách databáze savců. Pokud jej program nemá uložený v adresářové struktuře, dojde k jeho stažení.
- `WorldClimExtractor` - slouží ke stažení dat z klimatické databáze. Získaná data jsou souborového formátu ZIP. S využitím funkcionality třídy `Zip` jsou data extrahována a uložena do složky `worldClim2`. Zde jsou ve formátu TIFF připravena k dalšímu zpracování.
- `MatrixCreator` - transformuje získána data ze dvou databázi do výsledné Booleovské matice.

Důležité třídy balíčku `FactorFinder`:

- `FCF` - implementace metody GreConD, pracuje se vzniklou maticí a vrací seznam nalezených faktorových konceptů.
- `FactorPrep` - vytváří tři textové soubory `factorsInDetails.txt`, `averages.txt` `adifferences.txt`. Účel souborů popsán v hlavním textu práce. Slouží k interpretaci nalezených faktorových konceptů.
- `SvgCreator` - za pomocí SVG šablony vizualizuje faktor na mapě Evropy. Šablona `rendermap.svg` uložená ve složce `img/`, patří mezi kritické soubory. Šablona musí být přítomna při spuštění programu.

A.1 Spuštění programu

Před spuštěním programu je nutná instalace Java Runtime Environment 8 a vyšší. Vedle spustitelného souboru `extractor.jar` je vyžadována přítomnost souborů ve složce `downloadConfig/` a SVG šablony ve složce `img/`. Data z obou databází být přítomna nemusí. Defaultní verze na přiloženém DVD obsahuje tyto data spolu s výslednou maticí.

Program lze spustit za pomocí příkazové řádky: `java -jar extractor.jar`. Uživatel je průběžné informován o vykonávaných akcích, načtení/stažení dat, tvorba matice, hledání faktorových konceptů. Výstupem programu je matice uložena v řídkém formátu, soubor obsahující nalezené faktorové koncepty a soubory z něj odvozené. Posledním výstupem jsou vizualizace faktorových konceptů.

B Obsah přiloženého CD/DVD

`mammals/`
: Zde se nachází dva soubory, které vzniknou po úspěšné extrakci z databáze savců. Soubor `mammals.xml` obsahuje informace o savcích, konkrétně jména a výskyt savců. V souboru `CGRSJSON` jsou uloženy GPS koordináty pro lokace výskytu savců.

`matrix/`
: Uložený výsledný dataset. Soubor `colsHeaders.txt` obsahuje hodnoty ze záhlaví sloupců. V souboru `rowsHeaders.txt` jsou uloženy hodnoty ze záhlaví řádků, latinské názvy savců. Poslední soubor `data` obsahuje samotnou matici, uloženou v řídkém formátu.

`world2/`
: Výsledné soubory po extrakci z klimatické databáze. Data ve formátu obrázku TIFF jsou roztříděny do složek dle klimatických proměnných.

`img/`
: Obsahuje vizualizované faktorové koncepty. Obrázky jsou ve formátu SVG. Dále také obsahuje šablonu pro tvorbu uvedených vizualizací.

`src/`
: Veškeré zdrojové kódy. Zdrojový kód je rozdělen do dvou složek. První složka `ExtractionTool/` obsahuje zdrojové kódy související s extrakcí dat. Druhá složka `FactorFinder/` obsahuje zdrojový kód metody GreConD spolu s kódem vizualizace faktorů a kódem výpisu informací o faktorech do textových souborů.

`textOutputs/`
: Textová reprezentace nalezených faktorů. První zdejší soubor `factors-InDetails.txt` obsahuje hodnoty dosazené za indexy sloupců a řádků,

`averages.txt` pak zprůměrované klim. hodnoty na výskytech savců a průměrné klim. hodnoty faktoru. Soubor `differences.txt` zobrazuje odlišnost klim. hodnot na výskytu savců od průměrných klim. hodnot faktoru.

`readme.txt`
Instrukce pro spuštění programu spolu s popisem adresářové struktury.

U veškerých cizích převzatých materiálů obsažených na CD/DVD jejich zahrnutí dovolují podmínky pro jejich šíření nebo přiložený souhlas držitele copyrightu. Pro všechny použité (a citované) materiály, u kterých toto není splněno a nejsou tak obsaženy na CD/DVD, je uveden jejich zdroj (např. webová adresa) v bibliografii nebo textu práce nebo v souboru `readme.txt`.

Literatura

[1] MITCHELL-JONES, A. J.; BOGDANOWICZ, W.; KRYSTUFEK, B., et al. *The Atlas of European Mammals (Poyser Natural History)*. 1999. ISBN 0856611301.

[2] *WorldClim2*. Dostupný z: http://www.worldclim.org/.

[3] ZWILLINGER, Dan. *Crc standard mathematical tables and formulas*. 2018. ISBN 9781498777803.

[4] IVAN E. SUTHERLAND, Gary W. Hodgman. Reentrant polygon clipping. *Communications of the ACM*. 1974, vol. 17.

[5] *Line-Line Intersection.* Dostupný z: http://mathworld.wolfram.com/Line-LineIntersection.html.

[6] BĚLOHLÁVEK, Radim. Konceptuální svazy a formální konceptuální analýza. Dostupný také z: ⟨http://belohlavek.inf.upol.cz/publications/Bel_Ksfka.pdf⟩.

[7] BELOHLAVEK, Radim. INTRODUCTION TO FORMAL CONCEPT ANALYSIS. 2008. Dostupný také z: ⟨http://phoenix.inf.upol.cz/esf/ucebni/formal.pdf⟩.

[8] BELOHLAVEK, Radim; VYCHODIL, Vilem. Discovery of optimal factors in binary data via a novel method of matrix decomposition. *Journal of Computer and System Sciences*. 2010, vol. 76, s. 3–20.

[9] ARNFIELD, A. John. Köppen climate classification. 2009. Dostupný také z: ⟨https://www.britannica.com/science/Koppen-climate-classification⟩.

[10] GOSZCZYŃSKI J., Wasilewski M. Studies on the European hare. 46. Predation of foxes on a hare population in central Poland. *Białowieża*. 1992, vol. 37.

[11] *Wood Mouse (Apodemus sylvaticus).* https://www.mammal.org.uk/species-hub/full-species-hub/discover-mammals/species-wood-mouse/.

[12] PRESS, Johns Hopkins University. *Wild Mammals of North America: Biology, Management, and Conservation*. ISBN 9780801874161.

[13] HARRIS, Stephen K.; YALDEN, Derek William. *Mammals of the British Isles: handbook*. 2008.

[14] *Common Shrew – Sorex araneus.* https://www.mammal.org.uk/species-hub/full-species-hub/discover-mammals/species-common-shrew/.

[15] *Sorex minutus - Eurasian pygmy shrew.* Dostupný z: ⟨https://animaldiversity.org/accounts/Sorex_minutus/⟩.

[16] LANSZKI, Jozsef; ZALEWSKI, Andrzej; HORVÁTH, Győző. Comparison of Red Fox Vulpes Vulpes and Pine Marten Martes Martes Food Habits in a Deciduous Forest in Hungary. *Wildlife Biology*. 2007, s. 258–271.

[17] WEBER, J.-M; MEIA, J.-S; MEYER, Sandrine. Breeding success of the red fox Vulpes vulpes in relation to fluctuating prey in central Europe. *Wildlife Biology*. 1999, vol. 5, s. 241–244.
Dostupný také z: ⟨http://dx.doi.org/10.2981/wlb.1999.029⟩.

[18] *Lepus timidus - Mountain hare*.
Dostupný z: ⟨https://animaldiversity.org/accounts/Lepus_timidus/⟩.

[19] SHEEHY, Emma; LAWTON, Colin, 2015. Predators of red and grey squirrels in their natural and introduced ranges, s. 83–96. ISBN 978-0-9747576-1-8.

[20] *American Mink*.
Dostupný z: ⟨https://www.sciencedaily.com/terms/american_mink.htm⟩.

[21] BJØRN DAHLE, Kjell Wallin. Predation on adult moose Alces alces by European brown bears Ursus arctos. *Wildlife Biology*. 2013. Dostupný také z: ⟨https://doi.org/10.2981/10-113⟩.

[22] MORI, Emiliano; MENCHETTI, Mattia; BALESTRIERI, Alessandro. Interspecific den sharing: a study on European badger setts using camera traps. *acta ethologica*. 2014.
Dostupný také z: ⟨http://dx.doi.org/10.1007/s10211-014-0197-1⟩.

[23] KARLI, P. The Norway Rat's Killing Response To the White Mouse : an Experimental Analysis 1. *Behaviour*. 956, vol. 10, no. 1. Dostupný také z: ⟨https://doi.org/10.1163/156853956X00110⟩.

[24] UNWIN, Mike. *RSPB Spotlight: Foxes*. 2015.

[25] COLLINS. *Collins Wild Guide British Wildlife: The Essential Beginners Guide*. 2002. ISBN 0-00-713716-8.

[26] *Field Vole – Microtus agrestis*. https://www.mammal.org.uk/species-hub/full-species-hub/discover-mammals/species-field-vole/.

[27] WIKENROS CAMILLA Aronsson Malin, Liberg Olof. Fear or food – abundance of red fox in relation to occurrence of lynx and wol. *Scientific Reports*. 2017, vol. 7. Dostupný také z: ⟨https://doi.org/10.1038/s41598-017-08927-6⟩.