

Česká verze stránek není od roku 2013 aktualizována. Aktuální verzi této stránky najdete v její anglické verzi zde.

Import ekologických dat do programu R

Import z Excelu

Přes clipboard

Rychlá varianta, pokud potřebujete naimportovat data jen jednou a neplánujete je importovat znovu při každém použití skriptu. Otevřete soubor vltava.xls v Excelu. Použijte list *Vltava species*, který obsahuje matici druhy x vzorky. Celou matici (ovšem bez prázdných řádků nebo sloupečků) uložte do cliboardu (Ctrl+C). Pro načtení použijte funkci `read.delim`:

```
veg.data <- read.delim (file = 'clipboard', head = T, row.names = 1)
```

Funkce `read.delim` načte data z argumentu `file`, v tomto případě clipboardu. Funkce `read.delim` je odvozená z funkce `read.table` a je určená pro data oddělená tabulátory (`\tab`). Argument `head` značí, že se první řádek má považovat za názvy proměnných, `row.names` zase říká, že se první sloupec má použít pro názvy řádků.

Vzorová data v Excelu jsou ve formátu, který má druhy v řádcích a snímky ve sloupcích (z jednoduchého důvodu - druhů je víc než 256, a ve verzi Excelu 2003 není možné otevřít tabulku s víc jak 256 sloupci), a pro další použití v R je třeba matici transponovat:

```
veg.data <- t (veg.data)
```

Přes formát txt

Pokud plánujete uložená data načítat častěji (například při každém spuštění skriptu), je lepší, aby součástí skriptu byl i příkaz k vlastnímu načtení dat z externího souboru. Výhodou je použití jednoduchého formátu *.txt, ve kterém můžete data v Excelu uložit (vždy jen jeden pracovní list). Tento formát je výhodný proto, že jednotlivé buňky matice jsou vždy odděleny tabulátory (zatímco formát *.csv má různé varianty podle nastaveného jazykového prostředí systému). Pro vyzkoušení si stáhněte soubor vltava-spe.txt. Soubor buď uložte do pracovního adresáře eRka (ten zjistíme funkcí `getwd ()` a můžeme jej změnit funkcí `setwd ('c:/novy/adresar/')`), nebo do argumentu `file` v příkazu "read.delim" vložte celou cestu k souboru:

```
veg.data <- read.delim (file = 'c:/Dokumenty/vltava-spe.txt', row.names = 1)
```

Při vypisování celé cesty k adresáři je třeba používat normální lomítka (/), ne zpětná (\); pokud potřebujete použít lomítka zpětná, musíte je zdvojit: `c:\\novy\\adresar\\`. Data už není potřeba transponovat (tedy měnit sloupce za řádky), protože příkladový *.txt soubor už má druhy ve sloupcích.

Přes formát csv

Formát *.csv má nevýhodu v tom, že závisí na nastavení jazykové lokalizace na počítači. Pokud soubor exportujete z Excelu a máte anglickou lokalizaci, buňku se oddělí čárkami a jako oddělovače desetin se v číslech

použijí tečky. V české lokalizaci se ale jako oddělovače desetin tradičně používají čárky, a jako oddělovače buněk se proto používají středníky [1]. Další možností je data z Excelu uložit ve formátu `csv` (*comma separated values*) a eRku říct, jaký oddělovač byl použit pro oddělení jednotlivých buněk matice (středník nebo čárka). Použijeme například soubor vltava-env.csv (data oddělená středníky a desetiny oddělené čárkami). Soubor opět uložíme do pracovního adresáře eRka nebo do argumentu `file` vložíme celou cestu k souboru:

```
env.data <- read.table (file = 'c:/Dokumenty/vltava-env.csv', head = T, row.names =
1, sep = ';', dec = ',')
```

Argument ''sep = ';' '' označuje, že data jsou oddělená právě středníkem, argument ''dec = ',' '' ukazuje na čárku jako oddělovač desetinných míst.

Místo funkce `read.table` je v případě dat oddělených středníky možné použít funkci `read.csv2`, která má nastavený argument `sep = ';' '' a předpokládá, že desetiny jsou odděleny čárkami (tak jako v české lokalizaci Windows). Alternativní funkce ''read.csv` očekává jako oddělovače buněk čárky a jako oddělovače desetinných míst tečky.

```
env.data <- read.csv2 ('c:/Dokumenty/vltava-env.csv', row.names = 1)
```

Argument `file` může obsahovat i odkaz na umístění souboru na internetu:

```
env.data <- read.csv2
('http://www.davidzeleny.net/anadat-r/data-download/vltava-env.csv', row.names = 1)
```

Přímo z Excelovského souboru

Tato možnost je nejsložitější a navíc závislá na verzi Excelu, ve které máte data uložena. Existuje několik knihoven, které import z Excelu umožňují (přehled najdete například zde, ale tyto knihovny vyžadují ještě jiný nástroj než jen vlastní program R. Jako nejjednodušší se jeví použití knihovny `XLConnect`, která vyžaduje, aby počítač měl nainstalován také program Java (což většina počítačů tak i tak má). Nejdříve je tedy třeba knihovnu 'XLConnect' nainstalovat. Zkuste načíst první pracovní list z Excelovského souboru vltava.xls, který si stáhněte např. do adresáře `c:/Dokumenty`.

```
install.packages ('XLConnect')
library (XLConnect)
veg.data <- readWorksheetFromFile(file = 'c:/Dokumenty/vltava.xls', sheet = 1,
header = T, rownames = 1)
```

Funkce `readWorksheetFromFile` otevře soubor specifikovaný argumentem `file` a načte pracovní list, specifikovaný argumentem `sheet` (číslo nebo jméno listu). Další argumenty jsou obdobné jako v příkazu `read.table`.

Import z cc! formátu

Formát `cc!` (případně `cep`) se používá pro import dat do programu CANOCO 4.0, a je také jedním z formátů, jak se dají vegetační data exportovat z programu Turboveg for Windows (se kterým pracují vegetační ekologové). Pro jeho import do eRka slouží funkce `read.cep`, která je součástí knihovny `vegan`. Jako příklad můžeme použít soubor vltava.cep, který umístíme do pracovního adresáře a následně načteme:

```
library (vegan)
veg.data <- read.cep (file = 'vltava.cep', force = T)
```

Argument `force = T` je nutné nastavit jako souhlas, že si jste vědomi toho, že eRko může v průběhu načítání dat náhle spadnou a vy jste si tedy předtím uložili všechna data. Pro větší datové sobory (matice s více než 10 000 buňkami) je nutné nastavit ještě argument `maxdata` - maximální možný počet nulových hodnot (pokud víme velikost importované matice, můžeme nastavit hodnotu vypočtenou jako počet řádků x počet sloupců matice).

Import z JUICE

JUICE je program, který vegetační ekologové používají především k editaci vegetačních dat. Nejsnadnější způsob je použít funkci JUICE-R, tedy přímý export dat z JUICE do R. V programu JUICE otevřete soubor, který chcete exportovat, a v menu File > Export > Table > To R Project txt File (nebo klávesovou zkratku CTRL+W). Matice snímky x druhy se vyexportuje v souboru `table.txt`, který je oddělený tabulátory a najdete ho v podadresáři `bin/i386` v instalaci programu R. Načíst do eRka ho lze například takto:

```
JUICE.table <- read.table ('table.txt', sep = '\t', check.names = F, head = T, row.names = 1)
```

Nastavení argumentu `check.names = F` má dva důvody: funkci tak nevadí duplicitní hodnoty v názvech proměnných (druhů - v tabulce se tak můžou vyskytovat dva druhy se stejným jménem a případně vyskytující se i ve stejném vegetačním patře, i když to samozřejmě samo o sobě není dobře), a funkce se navíc nesnaží modifikovat názvy proměnných tak, aby byly syntakticky správně (například mezera v názvu by se nahradila tečkou).

[1)]
Pokud na počítači používáte program JUICE, pozor na jednu věc - JUICE občas změní nastavení počítače v české lokalizaci tak, aby používal desetinnou čárku místo tečky. Vzniká pak hybridní situace, kdy Excel ukládá do *.csv souboru data oddělená středníkem (jako v české lokalizaci), ale místo desetinných čárek používá tečky (jako v anglické lokalizaci).

From:
https://anadat-r.davidzeleny.net/ - **Analysis of community ecology data in R**

Permanent link:
https://anadat-r.davidzeleny.net/doku.php/cs:data_import

Last update: **2017/10/11 20:36**