Metadata for compilation of David Inouye's Gothic (Colorado, USA) long-term flowering data

Compiled by E. M. Wolkovich and J. Regetz,
with help from David Inouye, Amy Iler and Jane Ogilvie

17 September 2015

Contents

1 Basics

We started this project in 2010 because the original compilation by Jessica Forest did not include the following:

- Greenhouse plots, nor the stream plot
- Non-zoophilious species

So we (Lizzie Wolkovich & Jim Regetz) went about re-compiling the data from the original xls files that were edited by George Aldridge (1973-2009) and shared with Wolkovich. Note that at least one version of the data does contain these things (above), they may have been added in by George Aldridge but, either way, this recompilation contains more information than previous versions (more species and more years for some species).

2 Overview of re-compiling Excel files

The work of compiling these xls files occurs in two parts: a Perl script: `xls2csvJR.pl` and then an R script: `processraw.R`.

Data for each species were taken from the xls files. In general, there are two rows for each species name in the original files, with the first row *generally* representing a count of flowers and the second row *generally* representing a count of something else. Often 'what was counted' information is given in the last column of each spreadsheet. We extracted these data as follows:

1. If there were > 2 rows for a species, we excluded the data in the extra rows (usually there was only a third row, here are Regetz's notes from issue 44 in Redmine (project management software to manage the code and record issues):

   > Lizzie talked with George Aldridge Friday (Inouye postdoc) and it truly seems like the third column is either:
   > - the sum of the first two
   > - plants that look sorta male and sorta female
   > - a mystery, really, truly, who knows?
   > Based on this and that it's impossible to tell a mystery from a s/he plant we agreed strongly that tossing the third row of the data is the way to go.
   > Yee-haw!

2. If there were only 1 row, we caused an error to report (which we never saw)
3. If there were exactly two rows (most common case) we:
   (a) converted all NAs to 0
   (b) then converted all non-numeric values to NA
   (c) then compared row1 values to row2 values, excluding NAs and:
      i. if all $row1 >= row2$, we used row1 as 'flower' and row2 as 'other'
      ii. else if all $row2 >= row1$, do the reverse
      iii. else remove both rows of data.

Email from Jim from 2012 August plant parts (svn r48):

> In cases where the same plant part is given for both rows of data, the values from the first row were put into the appropriate column, and the values from the second row were put into a separate column called 'double.count'. In many cases this apparent duplication happens because one row actually represents males and the other row represents females, but we chose not to add the values together because we highly doubted that this was *always* the case (and it would have required throwing all the non-numeric values out, which appeared useful).
>
> Branches, catkins, stalks, and stems were included in 'other'. For any cases where the plant part column was empty (NA), the values were put into column 'unknown1' if they come from the first row for the species, or 'unknown2' if they come from the second row for the species. (Also, from taslk 127: In cases where no plant part column is detected, NA is used instead, which means the two rows of data for every species get inserted into the 'unknown1' and 'unknown2' columns.)

A few more notes:

- J. Forrest noted that there are some missing starts/ends of seasons and some months missing, of which it would be good to be aware.
- One datasheet issue: vr11984.xls and vr1-1984.xls appeared identical with the only difference appearing to be that the final column of the second row of the last species, "Senecio bigelovii": it contains "plants" in the former, but 'clumps(genets)' in the latter. We kept the 'clumps' one based on advice from G. Aldridge.
- All empty cells in the original Excel files are maintained in the main data file (`gothicclean.csv`), see below.

3 Plots

There was a lookup table for converting file names to plot names. G. Aldridge approved this.

`gothiccleanplots.R` does the work to pre-clean the plots I (Lizzie) then ran the output and decisions I made by George Aldridge (Inouye postdoc).

Plot naming:

- rm: rocky meadow
- gh: greenhouse
- vr: veratrum removal
- wm: wet meadow
- em: erthronium meados
- mw: meadow (alone) plots
- st: stream plots

- mi: interface plots (including willow-meadow interface plots)

Other changes:
numbers 10 and 11 – to 1
1973 plots were left unchanged on purpose (first year, naming conventions and plots in flux)
Note from Amy Iler on 10 September 2015: There is only one stream plot and one meadow plot (not to be confused with wet meadow plots), which were added in 2004 by David, with the aim of including more species. These need to be treated with caution when analyzing the data, because they can make it seem like there is a phenological advancement for species that were already present in the data set, when it is partly because these new plots were added. The greenhouse plots do not seem to have the same effect, probably because they were added earlier (e.g., analysis of change through time in first flowering date with and without the GH plots provides slopes that are correlated by $r = 0.98$).

4 Species name cleaning

`gothiccleansp.R` does the work to clean species names, a major undertaking by George Aldridge, David Inouye and Lizzie.

To clean the species names I (Lizzie) sent out the species list from the processed Excel files to George (and, towards the end, David) who then said which names were okay or not. I then added in all the changes and sent the new list out. Repeat. We repeated this process 6+ times (I think sometimes switching things back). I tried to coordinate all final information about these changes:

- `GothicInouye_phenspecies.csv`, should be a list of all species in the cleaned final file and relevant information, together:
- `cleanedspp.csv` and `cleaned_DIchanges.csv` *should, as best Lizzie can remember* contain a synposis of all species names and changes.

A couple more random notes:

- We treated all spp. as sp. based on email from G. Aldridge (dated 6.Apr.2011)
- From G. Aldridge 2010 March, "In the case of Valeriana, I believe flowers were never counted because they're so small, but there might be years where some enterprising soul did count them. There might also be times when the stalk counts for Valeriana are in row 1. I guess the way to know is if the numbers are really big or fairly small."
- At least two *Lupinus* species exist within the *Lupinus sp.* in the data. They generally have different phenologies. Ask Amy Iler, David Inouye or Jane Ogilvie for more information.

5 What's in the main data file?

The main data file is `gothicclean.csv` (currently in the species folder) – this is the final file after the merging and cleaning of Excel files, the cleaning of all the plot names and species

names.

See also: *Overview of re-compiling Excel files* above, which has many more details of what 'other' etc. actually means.

This file has 13 columns, here is Lizzie's 2015 understanding of them:

1. `species` – the cleaned species name
2. `plotNums` – the cleaned, shortened plot name and numbers (see section in this file on on Plots)
3. `filelow` – the original Excel file name in all lowercase
4. `file` – the original Excel file name, possibly not in all lowercase
5. `date` – the date of observation
6. `capitula` – number of capitula counted
7. `clusters` – number of clusters counted
8. `flowers` – number of flowers counted
9. `inflorescences` – number of inflorescences counted
10. `other` – number of other counted (includes branches, catkins, stalks, and stems)
11. `plants` – number of plants counted
12. `unknown1` – additional information that does not appear to be a capitula, clusters, flowers, inflorescences, other or plants (i.e., plant part column of original xls file was empty) and was from the *first* row of data for the species
13. `unknown2` – additional information that does not appear to be a capitula, clusters, flowers, inflorescences, other or plants (i.e., plant part column of original xls file was empty) and was from the *second* row of data for the species
14. `double.count` – Double count is in the `processraw.R` file and takes every instance where the species and plant part repeats on two rows of data and then takes whatever info there is in the second row and puts in the double-count column. Often this is a number, often it is a note such as 'tops of 2 stalks eaten' or 'frozen.') See also the 'email from Jim from August 2012' in *Overview of re-compiling Excel files.*
15. `plots` – plot in two letter code
16. `plotnums` – plot number (no letters)
17. `was.changed` – whether or not the species name was changed in species cleaning
18. `notes` – notes from the original file

6 What's in the folder 'gothicrecomp'?

A subset of all the files involved the production of the recompiling of the Inouye data, including all the above-mentioned `R` files and a `Perl` script (`xls2csvJR.pl`) that did the data cleaning, any csv files created or used to clean plot and species names, a few note files and:
`gothic-ambigous.csv`: As best I (Lizzie) can tell gothic-ambigous.csv is 43 times where the plant part was ambigously defined (as in one row it's called flowers and the other inflorescences or or there is a question mark after the plant part; or both are called inflorescences but have differing numbers – I looked some of these up and the few cases I looked at appear to be when male versus female were counted but I cannot say that all work that way). These data were included but should be checked if possible.

All the original Excel files we compiled from.

7 May also want to see

`gothic-testclean` is the folder received from Regetz in October 2010 and is the result of merging all the xls files from the `xls2csvJR.pl` and `processraw.R` files. It contains:

- `gothic-cleaned.csv` which Lizzie worked from for the plot and species cleaning. It also contains
- `gothic.Rout` this file contains lots of notes on the processing of the Excel files by Regetz. I (Lizzie) tried to go through many of these notes but I did not yet (I don't think) check the two cases with duplication mentioned:
  "Erigeron flagellaris?" in rm6-1981.xls
  "grass - long awns" in rm31991.xls
  If you run `processraw.R` (properly) this file should also be automatically re-generated.

`notes_on_compliling.txt` we never did use this but can include it and mention it if anyone else wants to. From Redmine:

> `notes_on_compiling.txt` file from J Forrest about which 'provides some information on whether flower or inflorescence data are in row 1 or 2 in which years "provides some information on whether flower or inflorescence data are in row 1 or 2 in which years" to see if we ended up generally picking these years, or how they were handled and what to do about it.

`Metadata file for David Inouye.doc` David Inouye's metadata file.

Note: The 'gothic' folder was version controlled through `Subversion` (on NCEAS' server, until 2015 when Lizzie stopped using `Subversion`) so there is a long log of files detailing when, by who and how all files were changed.

8 Acknowledgements

David Inouye, George Aldridge, Amy Iler, NSF grant DEB 0922080 (ask D. Inouye also),