Digitization Protocol

Sam Levin

2018-03-12

Contents

PADRINO

Plant IPM Database

MADRINA

Animal IPM Database

Padrino digitization protocol

Welcome to the `Padrino` and `Madrina` digitization team! The following is an in-depth guide for how to translate a published IPM into the *Excel* sheets and what to do with that sheet once you have completed the process.

There is a good chance I will have forgotten to include some important aspects of what this work entails. If that happens, please create an *Issue* in the GitHub repository as opposed to emailing me. This helps me (and, hopefully, you) keep all discussion related to the problem in a centralized location. Most importantly, this centralized location is accesible and searchable by your fellow digitizers, so they can also participate in the discussion or return to it later to reference it as needed.

This document assumes you have a general understanding of an IPM and why we want to have all of them in one place.

Data base structure

`Padrino` and `Madrina` are remotely hosted relational data bases that consider an individual IPM as the "atomic unit", as opposed to each vital rate or parameter. This helps keep relations tidy and avoid unnecessary data duplication, but can introduce some confusion at first. Hopefully, after reading this, the reasons for this decision will become more clear.

Relational data bases can be a bit tricky to work with if you aren't already familiar with them. My (admittedly limited) experience has been that this is not the most accesible format for most ecologists who are usually more accustomed to working in *R* or *Excel*. Therefore, we have created a "flat" version of it that is a set of 5 *Excel* sheets. The columns of each one are described below. The format of each description is: **data type**, description of variable, *constraint*.

Metadata

This is the table you will find when opening the spreadsheet. It contains important information about the study site, species, and authors that aren't necessarily relevant to construction of the IPM, but provide important context nonetheless.

ipm_id

Character

This column contains a unique identifier for each IPM in the data base. There should only be **1** row per IPM in this table.

This column cannot be blank.

species_author

Character

The genus and species epithet of the organism used by the author in the publication.

This column cannot be blank.

species_accepted

Character

The currently accepted genus and species epithet of the organism. For `Padrino` entries, this will come from the The Plant List. For `Madrina` entries, this will come from the Catalogue of Life.

This column cannot be blank.

tax_genus

Character

The accepted genus name of the organism. For `Padrino` entries, this will come from the The Plant List. For `Madrina` entries, this will come from the Catalogue of Life.

This column cannot be blank.

tax_family

Character

The accepted family of the organism. For `Padrino` entries, this will come from the The Plant List. For `Madrina` entries, this will come from the Catalogue of Life.

This column may be blank.

tax_order

Character

The accepted order of the organism. For `Padrino` entries, this will come from the The Plant List. For `Madrina` entries, this will come from the Catalogue of Life.

This column may be blank.

tax_class

Character

The accepted class of the organism. For `Padrino` entries, this will come from the The Plant List. For `Madrina` entries, this will come from the Catalogue of Life.

This column may be blank.

tax_phylum

Character

The accepted phylum of the organism. For `Padrino` entries, this will come from the The Plant List. For `Madrina` entries, this will come from the Catalogue of Life.

This column may be blank.

kingdom

Character

The accepted kingdom of the organism. For `Padrino` entries, this will come from the The Plant List. For `Madrina` entries, this will come from the Catalogue of Life.

This column may be blank.

organism_type

Character

`Padrino` entries

This is the general plant/algae type. This will usually come from the publication itself, but sometimes you may need to use other sources (e.g. other publications or taxonomic data bases) to find this information. Possible values are as follows

- Algae: brown, green or red. Green algae are in the *Plantae* kingdom, but are still considered algae for the purposes of this variable.
- Fungi: This includes fungus species, yeasts, molds, and multicellular fungi.
- Annual: This includes annuals and biennials. Annuals complete their entire lifecycle (birth, growth, reproduction, death) within a year wherease biennials can stretch that window to two years. For the sake of simplicity (and because both will die following reproduction), they are both classified as "Annual" in `Padrino`.
- Bryophyte: All bryophytes.
- Epiphyte: All epiphytes.
- Fern: All ferns species.
- Herbaceous perennial: All plants that herbaceous and have the potential to live for more than two years.
- Liana: All lianas.
- Palm: All palm species.
- Shrub: Woody upright plants that are not trees or palms.
- Succulent: All succulent species.
- Tree: All tree species.

`Madrina` entries

This is generally the same as `Class` for animals (except humans, which are recorded using their genus and species epithet). Non-animal species that are also not plants are typically recorded as *Bacteria* or *Virus*.

This column may be blank.

dicot_monocot

Character

Indicates whether a species is a dicot or monocot. Not applicable for `Madrina` entries.

This column may be blank.

angio_gymno

Character

Indicates whether a species is a angiosperm or gymnosperm. Not applicable for `Madrina` entries.

This column may be blank.

authors

Character

The last name of all authors. Multiple entries should be separated with semicolon (";").

This column cannot be blank.

journal

Character

The document that the information comes from. Possible values are listed below.

- Abbreviated name of the journal: We use the BIOSIS system for abbreviating journal names. More information on how to use it is in the link.
- Book: Models are sourced from a book.
- PhD Thesis: Models are sourced from a PhD thesis.
- MSc Thesis: Models are sourced from an MSc thesis.
- Report: Models are sourced from a report.
- Conference talk: Models are sourced from a conference talk.
- Conference poster: Models are sourced from a conference poster.

pub_year

Integer

The year that the model was published.

This column cannot be blank.

doi

Character

The *DOI* or *ISBN* for the publication.

This column may be blank.

corresponding_author

Character

The author to whom correspondance should be directed.

This column may be blank.

email_year

Character

The email address of the corresponding author with the year it is from in parentheses. If the email address is no longer in use, add the word "Dead" after a comma in the parentheses.

Example: levisc8@gmail.com (2018); levisc8@wfu.edu (2010, Dead)

This column may be blank.

remark

Character

Any observations you have about the model that are not captured by the other columns in `Metadata`.

This column may be blank.

apa_citation

Character

The full APA citation for the source.

This column may be blank.

demog_appendix_link

Character

If the model parameters are contained in an appendix, then include the link to said appendix here.

This column may be blank.

duration

Integer

Model duration is defined as the `end_year - start_year + 1`. Thus, this overlooks any years that were skipped.

This column cannot be blank.

start_year

Integer

The first year of data collection for the model.

This column cannot be blank.

start_month

Integer

The first month of data collection for the model. Months are numbered starting at 1 for January and continuing to 12 for December.

This column may be blank.

end_year

Integer

The last year of data collection for the model.

This column cannot be blank.

end_month

Integer

The last month of data collection for the model.

This column may be blank.

periodicity

Decimal

The number of times the model iterates per year. Periodic models with two transitions per year will have a value of two. Some IPMs for long lived tree species may have 5 year transitions; these will have a value of 0.2.

This column may not be blank.

number_populations

Integer

The number of populations described by the model.

This column may not be blank.

lat

Decimal

The decimal latitude coordinates for the site where data used in the model was collected from. Positive values refer to the northern hemisphere, while negative coordinates refer to the southern hemisphere.

This column may be blank.

lon

Decimal

The decimal longitude coordinates for the site where data used in the model was collected from. Positive values refer to the eastern hemisphere, while negative coordinates refer to the western hemisphere.

This column may be blank.

altitude

Decimal

The altitude in meters above sea level of the site where data used in the model was collected from.

This column may be blank.

country

Character

The ISO 3 country code.

This column may be blank.

continent

Character

The continent the study was conducted on. Possible values are below.

- n_america: Includes Canada, USA, and Mexico.
- s_america: Includes everything in the Americas except for Canada, USA, and Mexico.
- africa
- asia
- europe
- oceania: Various definitions exist, we are using this one.
- antarctica

This column may be blank.

ecoregion

Character

Indication of the ecoregion for the study, using the categories described in Figure 1 of Olson et al. (2001). For a more inclusive description of water ecoregions, see http://worldwildlife.org/biomes The one exception for this is the code LAB used for studies conducted in laboratory or greenhouse conditions. Possible values are below.

- TMB Terrestrial tropical and subtropical moist broadleaf forests
- TDB Terrestrial tropical and subtropical dry broadleaf forests
- TSC Terrestrial tropical and subtropical coniferous forests
- TBM Terrestrial temperate broadleaf and mixed forests
- TCF Terrestrial temperate coniferous forests
- BOR Terrestrial boreal forests/ taiga
- TGV Terrestrial tropical and subtropical grasslands, savannas and shrublands
- TGS Terrestrial temperate grasslands, savannas, and shrublands
- FGS Terrestrial flooded grasslands and savannas
- MON Terrestrial montane grasslands and shrublands
- TUN Terrestrial – tundra
- MED Terrestrial Mediterranean forests, woodlands and scrubs
- DES Terrestrial deserts and xeric shrublands
- MAN Terrestrial – mangroves
- LRE Freshwater large river ecosystems

- LRH Freshwater large river headwater ecosystems
- LRD Freshwater large river delta ecosystems
- SRE Freshwater small river ecosystems
- SLE Freshwater small lake ecosystems
- LLE Freshwater large lake ecosystems
- XBE Freshwater xeric basin ecosystems
- POE Marine polar ecosystems
- TSS Marine temperate shelf and seas ecosystems
- TEU Marine temperate upwellings
- TRU Marine tropical upwellings
- TRC Marine tropical coral
- LAB Laboratory or greenhouse conditions – controlled, usually indoor, conditions that mean the study species is not affected by the environment conditions typical of the actual geographic location of the study

This column may be blank.

studied_sex

Character

The sex of the individuals modeled.

- M Studied only males
- F Studied only females
- H Studied hermaphrodites
- M/F Males and females separately in the same IPM
- A All sexes modeled together

This column may be blank.

eviction_used

Boolean

Indicates whether authors corrected their discretized kernels for eviction (Williams et al. 2012). Possible values are *t* and *f*. *t* indicates eviction was corrected for, *f* indicates that it was not.

This column cannot be blank.

evict_type

Character

The type of correction used for eviction. Still working out the convention to use for this. . . .

This column may be blank.

treatment

Character

If a treatment was applied to the modeled population, indicate that here.

This column may be blank.

States

This table contains information on the state variables used by the authors to generate their IPM.

ipm_id

Character

This column contains a unique identifier for each IPM in the data base. However, this differs from the `Metadata` table in that a single IPM may have multiple rows in this table.

This column cannot be blank.

state_variable

Character

This column contains the name of a state variable used to construct an IPM *as reported by the authors of the paper*. For example, this could be **DBH** for a tree species or **Body_Mass** for an animal species. State variables do not need to be continuous. Examples of discrete state variables include reproductive status, age, or pathogen load (e.g. low, medium, high). A single model may use multiple state variables (and thus have multiple rows in this table, see above).

This column cannot be blank.

discrete

Boolean

This column indicates whether or not the variable is discrete or continuous. Use `"t"` to indicate a variable is discrete and `"f"` to indicate that it is continuous.

This column cannot be blank.

discrete_type

Character

This column indicates the type of discretization used for a discrete state variable. In depth explanations are provided in the Discrete Variables appendix.

This column may be blank.

Domains

This table contains information on the domains associated with each state variable in `States`. Keep in mind that one `state_variable` can have multiple domains (which themselves may be defined by a different `state_variable`!).

ipm_id

Character

This column contains a unique identifier for each IPM in the data base. However, this differs from the `Metadata` table in that a single IPM may have multiple rows in this table.

This column cannot be blank.

state_variable

Character

This column contains the same state variables as in the `States` table. However, a given `state_variable` may have multiple domains, so entries in this column need not be unique.

This column cannot be blank.

domain

Character

This column contains a unique identifier for the domain of each `state_variable`. For example, if **DBH** is implemented on 3 separate domains in the publication (perhaps for 3 different light environments in some megamatrix), then this could be named **size1**, **size2**, and **size3** for each one. The corresponding rows in `state_variable` should be filled in with **DBH** (i.e. not unique).

lower

decimal

In most cases, this will be a decimal or integer corresponding to the smallest value of the `state_variable` in the given domain.

upper

n_meshpoints

Integer

The number of bins that each domain is divided into for numerical integration. For some discrete variables, this will be blank.

This column may be blank.

Model Expressions (`ModelExpr`)

This table contains textual expressions of the models used to create IPM. A given IPM will have many rows in this table.

ipm_id

Character

This column contains a unique identifier for each IPM in the data base. However, this differs from the `Metadata` table in that a single IPM will have multiple rows in this table.

This column cannot be blank.

demographic_parameter

formula

See Writing Model Formulae for additional details

model_type

model_family

kernel_id

Model Values (`ModelValues`)

This contains the actual values for all of the parameters described in `ModelExpr`. A single IPM will have many rows in this table.

Every column in this table **must** be filled in to be able to enter the data base.

ipm_id

Character

This column contains a unique identifier for each IPM in the data base. However, this differs from the `Metadata` table in that a single IPM will have multiple rows in this table.

This column cannot be blank.

demographic_parameter

state_variable

parameter_type

parameter_name

parameter_value

Writing Model Formulae

Details on conventions and lots of examples!

Discrete Variables in Padrino and Madrina

The flexibility of IPMs to allow classification of individuals as a function of both continuous and discrete characters is one reason they are so powerful. Unfortunately, it also makes describing them in a data base a bit more complicated. Currently, there are a couple of supported types of discrete variables.

An IPM can be comprised of some number smaller subcomponents. For this data base, we describe each one separately in `ModelExpr` and then supply the `discrete_type` tag in the `States` table to indicate how those subcomponents need to be combined later on. I will attempt to describe each `discrete_type` in depth below.

- **Lefkovitch:** This `discrete_type` encompasses instances where authors have constructed a block megamatrix with components that are themselves IPMs. For example, Metcalf et al. (2009) generate a 6 x 6 megamatrix with each element corresponding to an IPM that describes transition probabilities from a given size and canopy illumination environment at time t to a givent size and canopy illumination environment at time *t+1*. Blocks on the diagonal of this matrix represent survival and growth transitions for trees that remain in the same light environment through time. Sub-diagonal elements represent survival and growth transitions for trees that move into more shaded environments, while super-diagonal represent survival and growth transitions for trees that move into lighter environments.

  In general, `discrete_types` of **Lefkovitch** will be used to represent variables where any transition between discrete states is possible.

- **Leslie/Age:** This `discrete_type` describes IPMs where the discrete variable of interest is ordered an individual **must** transition to the next the next value of the discrete variable if it survives through the projection interval. The most common example of this would be an age x size IPM (e.g. Childs et al. 2003). This can be represented as a block megamatrix where the top row of blocks represents age and size dependent fecundity and sub-diagonal blocks represent age and size depedent growth and survival. All other blocks are filled with 0s. This differs from the **Lefkovitch** designation because not all block-level transitions are possible (i.e. an individual cannot become 5 years old in at *t+1* if it is only 2 years old at t).

-