Production Workflow

06 August, 2017

Data Preparations

General from the Original Files (in `Data_Prepper.R`)

1. Load files in 'data/original/', prep as described in `Data_Prepper.R`, and output prepped files to 'data/prepped/'. Once properly prepped, `Data_Prepper'R` should not need to be run again. A simple log of each run of `Data_Prepper.R` is saved in 'data/prepped/dataPrepper_logs/'. These logs should be carefully examined each time new prepped files are made to compare processes to see if anything obviously went awry (i.e., items changed in the logs were as expected). See script for all decisions, but major decisions made here are described below.
    a. Population estimates (from `walleye_PE_data.csv`)
        i. Removed all WBICs that ended in a 1 (these are lake chains).
        ii. Found all unique WBIC_YEARs in this file as we need a PE to be able to compute P and B.
    b. WBIC characteristics file (from `ROW.csv` and `Supplementary Dataset 8 - Final Lake Class List.csv`)
        i. Joined the WBIC information from the DNR with Rypel's Data Supplement 8 that had lake classifications and other information in it.
        ii. A very small number of WBICs were from the "simple" lake classification types, so those were lumped as "simple."
    c. Weight-Length data (from `length_weight_age_raw_data_DATEHERE.csv`)
        i. Joined on lake classification, etc. information
        ii. Reduced to only those WBIC_YEARs with PEs.
        iii. Reduced to only fyke net caught fish from March, April, and May.
        iv. Removed fish for which the length (in) or weight (g) was less than 1. [*Legacy ... these fish no longer existed after previous decisions.*]
    d. Age-Length data (from `length_weight_age_raw_data_DATEHERE.csv`)
        i. Joined on lake classification, etc. information
        ii. Reduced to only those WBIC_YEARs with PEs.
        iii. Reduced to only fyke net caught fish from March, April, and May.
        iv. Removed fish less than 5 inches that were older than age 3.
        v. NOTE: found smallest age-3 fish for use in next step.
    e. FMDB lengths data (from `raw_data_walleye_DATE HERE.csv`)
        i. Joined on lake classification, etc. information
        ii. Reduced to only those WBIC_YEARs with PEs.
        iii. Reduced to only fyke net caught fish from March, April, and May.
        iv. Removed rows that were obvious errors (e.g., upper length greater than lower length, negative lengths).
        v. Generated individual lengths for those fish measured in length bins (*did this for inches data as that was the native recordings, then converted results to mm*).

Compute All Weight-Length Regressions (in `calcLWRegs.R`)

1. Load prepped `len_wt.csv` file.
2. Compute regressions and extract coefficients (loga and b), sample size (n), coefficient of determination (r^2), and range of lengths for (1) each WBIC_YEAR, (2) each WBIC, (3) each lake class, and (4) all fish regardless of any classification.
3. Combine all regressions into a single data.frame.

   a. Create a `use` variable that is set to `yes` if the regression is "valid" to use and `NO` if it is not valid to use. A regression was considered valid to use if it met the following criteria. **QUESTION: Did Goto assess the "validity" of weight-length regressions and, if so, how?**
      i. Sample size was above some minimum threshold. **Currently set at 25.**
      ii. r^2 was above some minimum threshold. **Currently set at 0.85.**
      iii. b was between 2 and 4 (Froese (2006) showed empirically that most MEAN (by species) values of b were between 2.5 and 3.5. Individual values would likely be a little wider; thus, the value set here.)
      iv. **QUESTION – Do we want to include a criterion based on the range of lengths in the regression? There are a handful of regressions with a fairly small range of values (~8 regression, of those remaining after above, with a range less than 200 mm).**
4. Output this data.frame to 'data/prepped/' as `LWRegs.csv`. This is used in `calcPB_loop` so that the regressions do not have to be run each time and so that the weight-length regression results can be more easily observed.
   a. Note that a script for simple exploratory summaries of the regression results is in the scratch folder.

Compute All Age-Length Keys (in `calcALKs.R`)

1. Load prepped `len_age.csv` file.
2. Attempt to compute empirical and smoothed ALKs by (1) each WBIC_YEAR, (2) each WBIC, (3) each lake class, and (4) all fish regardless of any classification.
3. Output each ALK to a ".RData" file that can be loaded later (thus, these are not created "on the fly" each time).
   a. Note that smoothed ALKs could not be reliably computed when n less than 10 or number of ages present was less than 3. Additionally, in a small number of situations, the model did not converge.
4. Created a `use` variable that is set to `yes` if the ALK is "valid" to use and `NO` if it is not valid to use. An ALK was considered valid to use if it met the following criteria. **QUESTION: Did Goto assess the "validity" of ALKs and, if so, how?**
   1. Sample size was above some minimum threshold. **Currently set at 30.**
   2. The number of ages in the ALK was above some minimum threshold. **Currently set at 5.**
   3. The number of length categories in the ALK was above some minimum threshold. **Currently set at 5 (i.e., using 20-mm length classes this ensure a range of at least 100 mm).**
   4. The smoothed ALK could not be fit (see above). It was assumed that if the smoothed ALK could not be fit that there were likely issues with the empirical ALK as well and, thus, both were deemed to not be "valid."
5. Output a data file (`ALKInfo.csv`) that contained information about each ALK. This is used in `calcPB_loop` so that the ALKs do not have to be constructed each time.
   a. Note that a script for simple exploratory summaries of the ALK results is in the scratch folder.

Computed P and B for each WBIC_Year (in `calcPB_loop.R`)

1. Loaded the prepped files (from data/prepped/ folder)
   a. Loaded the lake characteristics file (in `wbicInfo.csv`).
   b. Loaded the fmdb individual fish information file (in `fmdb_WAE.csv`).
   c. Loaded the PE values file (in `PE.cwv`).
      i. Removed WBIC_YEARs from the PE file that did not exist in the FMDB file (i.e., even if a PE existed, we can not calculate P without lengths to turn into weights and ages).
   d. Loaded the weight-length regression results file (in `LWRegs.csv`).
      i. Removed weight-length regressions that were deemed not valid (see above).
   e. Loaded the age-length key info file (in `ALKInfo.csv`).
      i. Removed ALKs that were deemed not valid (see above).

2. Looped through all remaining WBIC_YEARs in the PE file, computing P and B for each.
    a. The weight-length regression used was the first of these that was considered valid: WBIC_YEAR, WBIC, lake classification, all fish (no classifications).
        i. **QUESTION: Weight was estimated for individual fish and then the mean weight-at-age was calculated (see below). Did Goto do this or did he use the weight-length relationship to estimate mean weight from mean length.**
    b. The ALK used was the first of these that was considered valid: WBIC_YEAR, WBIC, lake classification, all fish (no classifications).
        i. **QUESTION: Ages were assigned to individual fish using the Isermann-Knight (2003) method. Did Goto do this or did he use "traditional" methods to estimate the proportions at age and mean length at age?**
        ii. **NOTE that we need to decide whether to use empirical or smoothed ALKs.**
    c. After age assignments, reduced to just age-3 fish (as the PEs were for just age-3 fish).
    d. Mean weight-at-age and number of fish sampled at age were summarized (from individual fish).
    e. The number of fish in the population at each age was determined by the proportion of the sample at each age and the PE. **QUESTION: What should we do when the sample size exceeded the PE?**
    f. These summaries were sent to `calcPB()` to compute P and B. [*This function has been validated in the sense that if it is given the same inputs as Rypel's Excel spreadsheet it computes the same P and B and intermediate results.*]
    g. An intermediate table (like Tables 1 in Rypel et al. manuscripts) was printed to an individual CSV (in results/CalcPB_Tables/) for each WBIC_YEAR for which a P could be calculated.
3. A summary table of P, B, other values (PE, n, number of ages, etc), and some notes was printed to a CSV file in the results folder following the loop. File is `PB_` with a suffix that is the date and time that the file was created (to allow comparisons if new decisions are made).
    a. Summary table included a `use` variable that is set to `yes` if the results met the following criterion for "validity."
        i. Sample size was above some minimum threshold. **Currently set at 30.** [*Goto used a threshold of 30 here.*]
        ii. The number of ages was above some minimum threshold. **Currently set at 5.** [*Goto used a threshold of 5 here.*]