README – BCGcalc R Tool (9/19/2018)

When applying the BCG, users should keep in mind that they can run any data through the model and get a result. However, ***if samples do not meet the criteria below, results should be interpreted with caution because they are outside the experience of the BCG model***.

Criteria

- **Size**: wadeable streams with drainage areas ranging from 1 to 100 mi^2
- **Geographic area**: Puget Lowlands and Willamette Valley EPA level3 ecoregions (numeric codes 2 & 3, respectively)
- **Stream type**: freshwater, perennial; does not cover unique habitats such as springs and seeps
- **Target number of organisms**: 500-count (subsampled to 600 total individuals where needed)
    - Note: a 300-count model is being developed as well
- **Sampling area**: ≥8 ft^2
- **Level of taxonomic resolution**: lowest practical level except for mites, which should be collapsed to the Order-level (Trombidiformes)
- **Collection gear**: D-Frame kick-nets with 500-micrometer net mesh
- **Collection method**: a targeted "riffle only" sampling scheme (like those used by King County and ODEQ) or WA ECY's multi-habitat, 'reach-wide' sampling scheme
- **Collection period**: July through October

Results should be interpreted with caution if they are flagged for any of the criteria listed in Appendix A (e.g., brackish influence, extreme dominance by one or two taxa).

Input file

The user must generate the input file to run through the BCGcalc R tool (which then calculates metric values, metric membership values, BCG level membership and BCG level assignments). The R code references the required fields (see Table 1) in the input file when making these calculations (note: when creating column headings in the input file, the user should utilize the naming scheme in the first column of Table 1). The BCG workgroup went through a lengthy process to reach consensus on the BCG attribute assignments (BCG_Attr) for the Puget Lowlands and Willamette Valley (Stamp and Gerritsen 2018). Several of the BCG rules are based on BCG attribute metrics. Thus, for the BCG model results to be accurate and valid, users should make sure the BCG attributes in their input file match with those that are in the Excel file titled 'TaxaMaster_Bug_BCG_PacNW_v1.' If the input file contains taxa that are not on this list, the user should enter the appropriate phylogenetic information and a 'NA' in the 'BCG_Attr' field.

Metric calculations

The BCGcalc R tool can be used to calculate a multitude of metrics (beyond the 12 that are used in the BCG model), including thermal indicator, FFG, habit, tolerance value and life cycle metrics. The BCG workgroup went through a lengthy process to reach consensus on the thermal indicator designations for the Puget Lowlands and Willamette Valley (Stamp and Gerritsen 2018) but did not address the other attributes at a regional scale. If the FFG, habit, tolerance value and life cycle attribute fields are utilized, those designations need to come from the user. Users can calculate the full suite of metrics or adapt the code so that the output only includes a subset of metrics (for example, you may want to limit the output to only the 12 metrics used in the BCG model). Examples of R code that can be used to pull subsets of metrics can be found in the 'Saving Specific Metrics' section of the vignette.

Key files –

These are automatically uploaded to the 'extdata' folder when you install the R package (on my computer, they are located here: C:\Programs\R\R-3.4.1\library\BCGcalc\extdata). Of these, only the 'Rules' file is referenced by the R code.

- **Rules** (Excel) - contains worksheets with the BCG rules that the R code references; the worksheet titled 'BCG_PacNW_v1_500ct' is used for the 500-count model; the worksheet titled 'BCG_PacNW_v1_300ct' is used for the 300-count model.
- **vignette_BCGcalc** (html) – covers the basics going from raw data to BCG model results.
- **TaxaMaster_Bug_BCG_PacNW_v1** (Excel) – this file helps users create their input data files. It contains the taxa list from the Puget Lowlands and Willamette Valley BCG project, and associated attribute information (NonTarget, BCG attribute, Thermal_Indicator, FFG, Habit, LifeCycle, TolVal). Important notes:
    - When users create input files, they should make sure the **BCG attribute assignments** match with the designations in this file.
    - **Thermal_Indicator** assignments were based on temperature tolerance analyses and expert elicitation for the Puget Lowlands and Willamette Valley.
        - For detailed information on how these designations were made, see the Excel file titled 'PL_WV_ThermalIndicator_20180326'
    - **Functional feeding group (FFG)** assignments were compiled from ODEQ, WA ECY and the EPA National Aquatic Resource Surveys (NARS) (using majority rules; the BCG work group did not discuss or try to reach consensus on these)
    - The **Habit, LifeCycle & TolVal** fields (highlighted in red) have placeholder entries (the BCG work group did not try to reach consensus on these; if a user wants to generate metrics based on these fields, they will need to use their own data to populate these fields)
- **MetricNames** (Excel) – contains a list of all the potential metrics that can be calculated with the BCGcalc R tool.
- **Data_BCG_PacNW (**Excel) – example input file (it contains data for the 678 samples that were in the BCG calibration dataset). Users can use it as a template when preparing their own input data. Column headings for required fields are highlighted in green.

Disclaimer

Version 1 of the BCG model has known limitations and will need further testing in coming years. The models should be regarded as beta versions, with potential for refinement over time as the models are used with new data.

Literature cited

Stamp, J. and J. Gerritsen. 2018. Calibration of the Biological Condition Gradient (BCG) for Macroinvertebrate Assemblages in Puget Lowland/Willamette Valley Freshwater Wadeable Streams. Prepared by Tetra Tech for the US EPA Office of Water, Office of Science and Technology and US EPA Region 10

Table 1. List of required fields for the input file, with descriptions.

Required field	Description	Acceptable entries	Example entry	Notes
SampleID	Unique sample identifier	characters or numbers, must be unique	00004CSR_Bug_ 2000-06-28_0	The user can use whatever sample identifier scheme they typically use (in this example we used StationID_Assemblage_CollDate_Rep).
Index_Name	Name of the BCG rules worksheet that the R code is referencing	'BCG_PacNW_v1_500ct'	BCG_PacNW_v1 _500ct	***Must match with the names of the worksheets in the Rules Excel file***.
Site_Type	The user needs to select which model to apply: Low (Lo) or high (Hi) gradient. Low <1% NHDv2 flowline slope; high ≥ 1% NHD v2 flowline slope	'Lo' or 'Hi'	Lo	The designations in the BCG calibration dataset were based on the NHDPlus v2 flowline slope. The 1% threshold should be regarded as a fuzzy versus distinct line, as it represents a transitional zone where streams likely share characteristics of both low and high gradient stream types. Because of this, ***we recommend that users run sites with 0.5% to 1.5% slope through both BCG models and report both sets of results***. Users can also base their gradient designations on other sources (such as reach-scale measurements).
Area_mi2	Drainage area (mi^2), based on exact watershed delineations	Number; if not available, leave blank (which R will read as NA).	1.10	The BCG model is calibrated for ***wadeable sites with drainage areas ranging from 1 to 100 mi2***.
SurfaceArea	Sampling area in ft^2	Number; if not available, enter 'NA'	8	BCG model is calibrated for samples with ***sampling areas ≥ 8 ft2***
TaxaID	Unique taxa identifier	character or number, must be unique	Brillia	Taxa should be identified to the ***lowest practical taxonomic resolution except for mites, which should be collapsed to the Order-level (Trombidiformes)***. If this schema is not used, results should be interpreted with caution.
N_Taxa	Number of individuals	number	14	The BCG model is calibrated for ***500-organism samples (subsampled to 600 total individuals where needed)***

Table 1 continued…

Required field	Description	Acceptable entries	Example entry	Notes
Exclude	Redundant taxa are excluded from richness metric calculations but are counted in the other metrics	TRUE or FALSE (redundant taxa should be entered as "TRUE")	FALSE	For more information on how 'Exclude' designations were made for BCG model calibration, see Appendix B
NonTarget	Non-target taxa are not part of the intended capture list; e.g., fish, herps, water column taxa. They are excluded from all metric calculations	TRUE or FALSE (NonTarget taxa should be entered as "TRUE")	TRUE	NonTarget designations used in the BCG calibration dataset can be found in the 'TaxaMaster_Bug_BCG_PacNW_v1' Excel file. These designations are consistent with those used for WA & OR's indices (B-IBI & O/E, respectively)
Phylum	Phylogeny	Text; if not available, leave blank	Arthropoda	source file: TaxaMaster_Bug_BCG_PacNW_v1
SubPhylum			Hexapoda	
Class			Insecta	
SubClass			Pterygota	
Order			Ephemeroptera	
Family			Baetidae	
SubFamily				
Tribe				
Genus			Baetis	
SubGenus				
Species			tricaudatus complex	
BCG_Attr	BCG attribute assignment for the Puget Lowlands and Willamette Valley (Stamp and Gerritsen 2018)	1i, 1m, 2, 3, 4, 5, 6; if not available, leave blank or enter 'NA'	3	Source (Excel) file: TaxaMaster_Bug_BCG_PacNW_v1. If you use different BCG attribute assignments than this, the BCG model output may not be accurate

Table 1 continued…

Optional field	Description	Acceptable entries	Example entry	Notes
Thermal_Indicator	Thermal indicator designations for the Puget Lowlands and Willamette Valley. For information on how these designations were derived, see Excel file 'PL_WV_ThermalIndicator_20180326'	cold, cold_cool, cool_warm, warm	cold_cool	Source (Excel) file: TaxaMaster_Bug_BCG_PacNW_v1
FFG	Functional Feeding Groups: collector-gatherer (CG), collector-filterer (CF), predator (PR), scraper (SC), shredder (SH)	CG, CF, PR, SC, SH	PR	Designations need to come from the user (we did not attempt to reach regional consensus/reconcile differences across entities)
Habit	Habit: burrowers (BU), climbers (CB), clingers (CN), sprawlers (SP), swimmers (SW)	BU, CB, CN, SP, SW	CN	
Life_Cycle	univoltine (UNI), semivoltine (SEMI), multivoltine (MULTI)	UNI, SEMI, MULTI	UNI	
TolVal	0 to 10 scale, with 0 representing the tolerance value of an extremely sensitive organism and 10 for a tolerant organism	number	7	

Appendix A – Flagging criteria

Table 1. Samples are flagged for further evaluation if certain criteria are not met. BCG model outputs for flagged samples should be interpreted with caution.

Flags	Criteria	Rationale
Too few organisms	Raw counts < 450 total individuals in 500-count samples	The 450-individual threshold allows for a 50-organism ‘buffer’ in case the taxonomist rejects some organisms that the sorting technician includes. Something odd is going on if you don't get at least 450 organisms.
	Density < 500 organisms/m2 (or < 47 organisms/ft2)	Density is not available for all sites, otherwise this metric may be preferred over raw counts because it takes into account differences in sampling area and subsampling effort. Moving ahead, we recommend that density be included as a standard output and that this threshold be further explored and refined.
Too many organisms	> 600 total individuals in 500-count samples	Outside the bounds of experience of the model
Dominance of one or two taxa	% Individuals – most dominant two taxa ≥ 50%	Resampling should be performed if resources permit. Dominance of one or two taxa affects richness metrics in particular. There are no easy answers on how to address it (if labs kept picking beyond 500 organisms, they’d likely get more of the same - e.g., 600 Baetis tricaudatus instead of 300). If this occurs in BCG level 3 samples, panelists suspect that sensitive taxa are present but are being missed due to the dominant taxa; as sites get more degraded, panelists don’t expect as many taxa are being missed (they expect more of the same tolerant taxa). The dominance issue accounts for a fair amount of year-to-year variability (which affects IBI scores as well as BCG scores).
Size/drainage area	Very small (< 2 mi^2 drainage area)	User should be aware that very small streams have a higher likelihood of going dry in certain years and may have a higher variation in insect abundance
	Too large (> 100 mi2 drainage area)	Outside the bounds of experience of the model
Reduced sampling effort	Sampling area ≥ 8 ft2	Higher likelihood of collecting a poor sample and missing sensitive taxa
Brackish	Americorophium or Gnorimosphaeroma are present	Brackish water will affect the freshwater community. Note – Ramellogammarus (may be ID’d as Anisogammaridae or Eogammarus in some data) is estuarine, but it does often penetrate upstream into pure fresh water. Mysis or Mysidae may also indicate brackish water (but these taxa do not occur in our BCG calibration dataset)
Unique	Number of Ramellogammarus individuals ≥ 10	Abundances are highly variable (in the BCG dataset, number of individuals ranged from 28-229 at the same site). The threshold of 10 is an arbitrary starting point. We are not yet sure what high numbers of Ramellogammarus indicate.

Appendix B – Excluded Taxa Decision Criteria

When calculating metrics for benthic macroinvertebrates there are occasions when certain taxa are not included in taxa richness metrics but the individuals are included for all other metrics. This is done to avoid double counting taxa that may have been identified to a more coarse level when taxa of a finer level are present in the same sample.

These taxa have been referred to by many names – e.g., Excluded Taxa, NonUnique Taxa, or Ambiguous Taxa. This document will use the term Excluded. This is done so that taxa to be removed from taxa richness calculations have to “opt in”. That is, it is assumed all taxa will be counted. Only those taxa marked as “Excluded” will be removed.

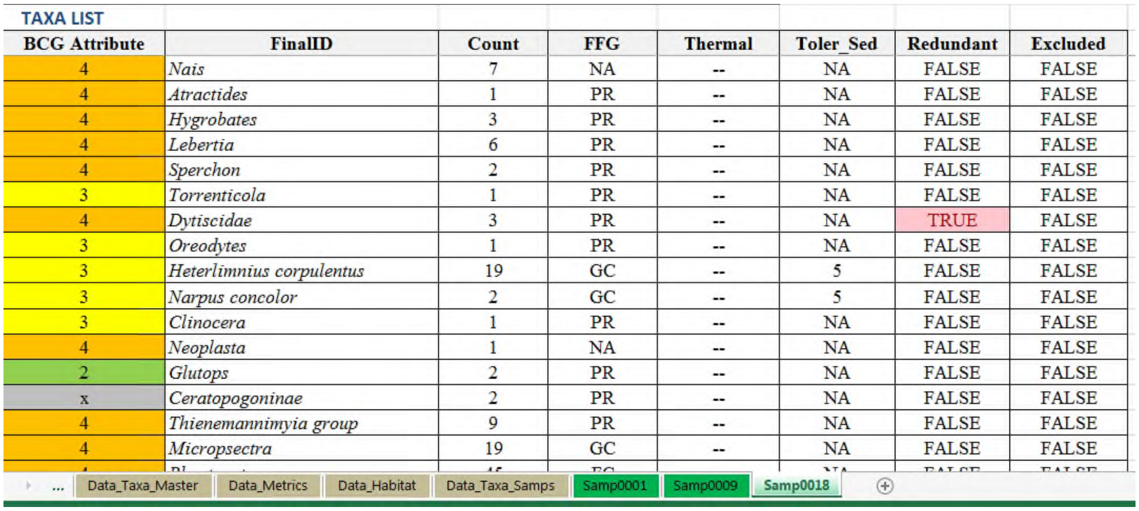
When version 1 of the BCG model was calibrated, redundant taxa were excluded from richness metric calculations based on the following steps:

1. Calculate and find all taxa names that appear in a sample at each taxonomic rank more than once (for an example, see Figure 1). These are the potential "parents" to be excluded.

2. Check if any of the potential "parents" equal a final ID in their respective samples.

3. If you get a match these were marked as "Excluded"

All Excluded decisions were sample-specific and the rules were reapplied if sample contents changed. Also, if the level of effort or operational taxonomic units changed, the Excluded taxa designations were recalculated.



TAXA LIST

BCG Attribute	FinalID	Count	FFG	Thermal	Toler_Sed	Redundant	Excluded
4	*Nais*	7	NA	--	NA	FALSE	FALSE
4	*Atractides*	1	PR	--	NA	FALSE	FALSE
4	*Hygrobates*	3	PR	--	NA	FALSE	FALSE
4	*Lebertia*	6	PR	--	NA	FALSE	FALSE
4	*Sperchon*	2	PR	--	NA	FALSE	FALSE
3	*Torrenticola*	1	PR	--	NA	FALSE	FALSE
4	*Dytiscidae*	3	PR	--	NA	TRUE	FALSE
3	*Oreodytes*	1	PR	--	NA	FALSE	FALSE
3	*Heterlimnius corpulentus*	19	GC	--	5	FALSE	FALSE
3	*Narpus concolor*	2	GC	--	5	FALSE	FALSE
3	*Clinocera*	1	PR	--	NA	FALSE	FALSE
4	*Neoplasta*	1	NA	--	NA	FALSE	FALSE
2	*Glutops*	2	PR	--	NA	FALSE	FALSE
x	*Ceratopogoninae*	2	PR	--	NA	FALSE	FALSE
4	*Thienemannimyia group*	9	PR	--	NA	FALSE	FALSE
4	*Micropsectra*	19	GC	--	NA	FALSE	FALSE

Figure 1. Example - Dytiscidae (family-level) is excluded from the richness metrics in this sample because these organisms could be the same taxon as Oreodytes (genus-level). The exclusion rule is applied on a sample by sample basis.

Below is a more detailed description of the process that was followed during BCG model calibration. Before starting it is necessary to have a complete and correct master taxa list (all phylogenetic information and ranks).

Terminology

- Target Rank = intended level of taxonomy for identification, e.g., genus. Typically, specified in the project's SOP but can be adjusted during the OTU process.
- Parent or Parent Taxon = a taxon that occurs in the data in addition to other taxa in the same group that are identified to a more specific level. For example, the family Baetidae may occur in the data in addition to genera within the family Baetidae. In this case the name Baetidae is a parent to the other taxa within the family. Parents do not have to be only a single rank above the child taxon. That is, the class and order ranks are parents of any family ranks within them.
- Child or Children Taxa = a taxa or taxon that occurs in the data in addition to individuals identified to a coarser level. For example, the genera Baetis and Procloeon may occur in addition to the family Baetidae (of which the 2 genera listed are a member). In this case Baetis and Procloen are children of Baetidae.

Rule Development

For each sample:

1. Determine "potential" taxa for exclusion based on rank (or level) names appearing more than once in a sample.
    a. This is done for all ranks present; phylum, class, order, family, tribe, genus, species.
2. Check if any "potential" taxa are equal to a final (unique) ID in the same sample.
3. Stage is combined with taxa names if used in the dataset.

Requirements

1. A sample taxa table or data frame.
    a. All non-count and zero individual taxa have been removed.
    b. Unique sample ID code in a single column.
    c. A column with a final identification that is narrative not numeric. That is, Baetidae is ok but the IT IS number is not.
    d. Phylogenetic rank/level columns.
        i. This can be applied from a master taxa table but needs to be included in this table. One column per rank.
        ii. Names need to be consistently spelled.

Procedures

1. Find all potential Parents (those with a rank coarser than the target rank). This is done by creating a list of taxa rank names that appear more than once in a sample. This is done for each taxonomic rank.
2. The above list is compared to the final identifications for each sample.
    a. Special consideration is made for ranks of finer detail than genus. That is, names that are a combination of more than one field.
3. Any matches are marked as "Excluded".

There is still a need for manual review / QC check of the final list of Excluded designations.