Supporting Information

Use and misuse of trait imputation in ecology, the problem of using out-of-context imputed values.

Lucas D. Gorné[1, 2,*], Jesús Aguirre-Gutiérrez[3, 4], Fernanda C. Souza [5], Nathan G. Swenson[6], Nathan Jared Boardman Kraft[7], Beatriz Schwantes Marimon[8], Tim R. Baker[9], Renato A. Ferreira de Lima[10], Emilio Vilanova[11], Esteban Alvarez-Davila[12], Abel Monteagudo Mendoza[13, 14], Gerardo Rafael Flores Llampazo[15], Rubens Manoel dos Santos[16], Gerhard Boenisch[17], Alejandro Araujo-Murakami[18], Gonzalo Francisco Rivas Torres[19], Hirma Ramírez-Angulo[20], Nayane Cristina dos Santos Prestes[21], Paulo S. Morandi[22], Sabina Cerruto Ribeiro[23], Wesley Jonatar Cruz[24], Mathias Disney[25, 26], Anthony Di Fiore[27, 28], Ben Hur Marimon-Junior[8], Ted R. Feldpausch[29], Yadvinder Malhi[3], Oliver Phillips[30], David Galbraith[30], Sandra Díaz[1,2].

Corresponding author (*)

[1]Universidad Nacional de Córdoba, Facultad de Ciencias Exactas Físicas y Naturales. Córdoba, Argentina.
[2]Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, IMBiV. Córdoba, Argentina.
[3]Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford OX13QY, UK.
[4]Leverhulme Centre for Nature Recovery, University of Oxford, Oxford OX13QY, UK.
[5]Departamento de Ecologia e Conservação, Instituto de Ciências Naturais, Universidade Federal de Lavras, Lavras, Minas Gerais, Brazil.
[6]Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 46556, USA.
[7]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, 621 Charles E. Young Drive South, Los Angeles, CA 90095, USA.
[8]Universidade do Estado de Mato Grosso (UNEMAT), Programa de Pós-Graduação em Ecologia e Conservação, Nova Xavantina, MT, Brazil.
[9]School of Geography, University of Leeds, UK.
[10]Departamento de Ciências Biológicas, ESALQ, Universidade de São Paulo, Avenida Pádua Dias 11, 13418-900 Piracicaba, Brazil.
[11]Wildlife Conservation Society, 2300 Southern Blvd Trail, Bronx, NY 10460, USA.
[12]UNAD-Universidad Nacional Abierta y a Distancia de Colombia, Colombia.
[13]Universidad Nacional de San Antonio Abad del Cusco, Perú.
[14]Jardín Botánico de Missouri, Perú.
[15]Instituto de Investigaciones de la Amazonía Peruana, Iquitos, Perú.
[16]Laboratory of Phytogeography and Evolutionary Ecology, Universidade Federal de Lavras, Brazil.
[17]Max-Planck-Institute for Biogeochemistry, Hans-Knoell-Str. 10, 07745 Jena, Germany.
[18]Museo de Historia Natural Noel Kempff Mercado, Universidad Autónoma Gabriel Rene Moreno, Bolivia.
[19]Universidad San Francisco de Quito-USFQ, Ecuador.
[20]Indefor, Facultad de Ciencias Forestales y Ambientales, Universidad de Los Andes, Mérida, Venezuela.
[21]Universidade do Estado de Mato Grosso, Brazil.
[22]Programa de Pós-graduação em Ecologia e Conservação, Universidade do Estado de Mato Grosso, Brazil.
[23]Centro de Ciências Biológicas e da Natureza, Universidade Federal do Acre, Brazil.
[24]AMAP (botAnique et Modélisation de l'Architecture des Plantes et des Végétations), 8 CIRAD, CNRS, INRAE, IRD, Montpellier Cedex 5, France.
[25]University College London, Dept of Geography, Gower Street, London, WC1E 6BT, UK.
[26]NERC National Centre for Earth Observation (NCEO), Gower Street, London, WC1E 6BT, UK.
[27]Department of Anthropology and Primate Molecular Ecology and Evolution Laboratory, The University of Texas at Austin, USA.
[28]Tiputini Biodiversity Station, Universidad San Francisco de Quito, Ecuador.
[29]Geography, Faculty of Environment, Science and Economy, University of Exeter, Exeter, UK.
[30]School of Geography, University of Leeds, Leeds, UK.

Table S1. number of observations for each trait in each one of the used datasets.

Dataset	total N° of entities	H	SSD	LA	LMA	N_{mass}
Baraloto species-level	448	442	439	448	448	403
Baraloto individual-level	7227	4255	3156	7148	7138	2830

Table S2. Ranges of variation of empirically measured traits and whether imputed values fall out of them in the species level Baraloto's dataset. H: adult plant height; SSD: stem specific density; LA: leaf area, LMA: leaf mass per area; N_{mass}: mass-based leaf nitrogen content.

Trait	Empirical range	Imputed below lower limit	Imputed over upper limit
H	4.98 – 48 m	0.113 m	(48.01, 91.70) m
SSD	0.311 – 0.905 g/cm^3	none	none
LA	9.77 – 108698.33 mm^2	none	none
LMA	40.56 – 255.83 mg/cm^2	none	none
N_{mass}	11.12 – 42.71 mg/g	(5.08, 5.67) mg/g	(64.80, 117.36, 104.68, 368.00) mg/g

Table S3. Summary tables for the analyses of the relative error of the imputed values in the individual-level Baraloto's dataset. Model: $\log_{10}(|\text{relative error}|+1) \sim \log_{10}(\text{actual value}) + \text{within-species relative range amplitud}$.

trait	coefficient	estimate	S.E.	t-value	p-value
H	intercept	1.772	0.045	39.06	<0.0001
	$\log_{10}$(actual value)	-0.530	0.037	-14.46	<0.0001
	relative range	0.429	0.041	10.54	<0.0001
SSD	intercept	0.513	0.020	26.17	<0.0001
	$\log_{10}$(actual value)	-1.118	0.071	-15.74	<0.0001
	relative range	0.493	0.045	11.01	<0.0001
LA	intercept	1.973	0.055	35.75	<0.0001
	$\log_{10}$(actual value)	-0.167	0.015	-11.03	<0.0001
	relative range	0.454	0.046	9.88	<0.0001
LMA	intercept	1.817	0.063	28.76	<0.0001
	$\log_{10}$(actual value)	-0.409	0.033	-12.35	<0.0001
	relative range	0.580	0.042	13.80	<0.0001
N_{mass}	intercept	1.079	0.085	12.621	<0.0001
	$\log_{10}$(actual value)	-0.232	0.068	-3.407	<0.0001
	relative range	0.851	0.085	10.00	<0.0001

Figure S1.

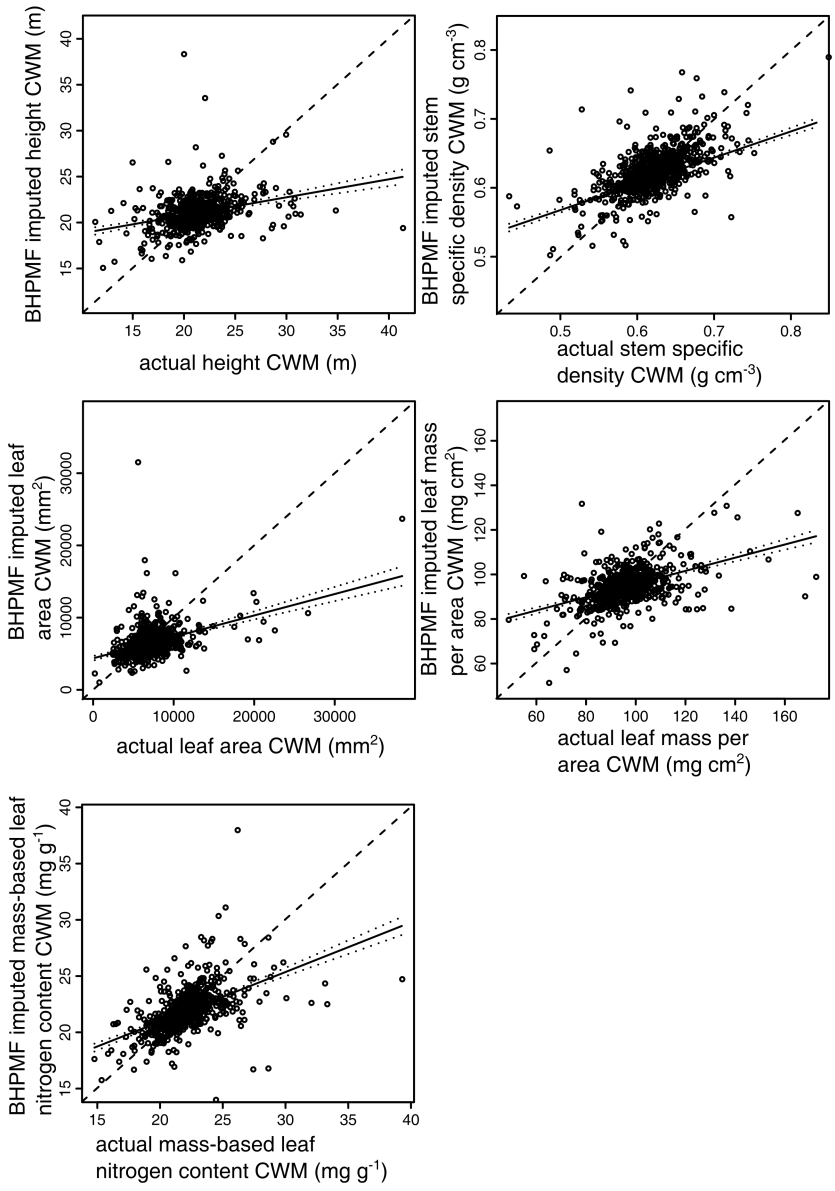

Figure S1. Relationship between community weighted means for each trait in the species-level Baraloto's dataset computed, for simulated communities, with the actual value of each trait vs the imputed values by the BHPMF method. The dashed line represents the identity line (intercept=0; slope=1). The continuous and dotted lines represent the fitted ordinary least-squares linear regression line and its 95% confidence interval.

Script code 1. Script performing the assessing the accuracy and precision of the imputation.

```
#download the version 0.3.3 of purrr from
https://cran.r-project.org/src/contrib/Archive/purrr/


#Running R 3.4.4

#required packages

install.packages("magrittr")

install.packages("rlang")


install.packages(".../purrr_0.3.3.tar.gz", # directory of the downloaded file

                    repos = NULL,

                    type = c("source"),

                    INSTALL_opts = "--no-test-load",

                    lib = "...") #set the path

install.packages("devtools")

install_github("fisw10/BHPMF")


library(devtools)

library(BHPMF)


#loading data

setwd("")

datos <- read.csv("datos.csv")

#It requires a data frame of entities (g0) by traits, plus the grouping of taxonomic
factors.

#The entities might be individuals, species, or other taxa.

#If the entities are individuals, the grouping factors are species, genus, and family.

#If the entities are species, the grouping factors are genus, family, and order.

#In this example the grouping factors will be "g1", "g2", "g3". The traits will be "T1" to
"T6".


#setting temporary directory for

tmp.dir <- dirname(".../tmp/") #"..." is the working directory


#prepearing needed objects
```

```
#trait.info must be a row-nameless matrix, with only names in the columns

# the rows order must be the same as in the hierarchy.info matrix

trait.info <- trait.raw <- as.matrix(datos[order(datos$g0), c("T1", "T2", "T3","T4", "T5",
"T6")])


#hierarchy.info cannot contain NAs. It must have taxonomic levels from less to more
comprehensive.

#hierarchy.info is a data.frame

hierarchy.info <- datos[order(datos$g0), c("g0", "g1", "g2", "g3")]


#checking if all individuals in the trait matrix are accounted for in the hierarchy matrix

nrow(hierarchy.info) == nrow(trait.info) #TRUE


#The traits must have a frequency distribution approx to the normal, and they must be
standardized.

#usually log10 y z-transformed

trans_par <- data.frame(trait=c("T1", "T2", "T3", "T4", "T5", "T6"), minx=NA, mlogx=NA,
slogx=NA)


i=1

for(i in 1:ncol(trait.info)){

       x <- trait.info[,i] # goes through the columns


       min_x <- min(x,na.rm = T) # takes the min of each column


       if(min_x < 0.00000000001){

              x <- x - min_x + 1 # make this optional if min x is neg

       }

       logx <- log10(x)

       mlogx <- mean(logx, na.rm = T)

       slogx <- sd(logx, na.rm = T)

       x <- (logx - mlogx)/slogx # Z transformation

       trait.info[,i] <- x


       trans_par$minx[i] <- min_x

       trans_par$mlogx[i] <- mlogx

       trans_par$slogx[i] <- slogx
```

```
}
trans_par
#trait       minx      mlogx      slogx
#   T1
#   T2
#   T3
#   T4
#   T5
#   T6


#Only to know and report the imputed values out of the empirical range
set.seed(1234) #always the same random seed
GapFilling(trait.info, hierarchy.info,
       prediction.level = 4,
       used.num.hierarchy.levels = 3,
       mean.gap.filled.output.path = paste0(tmp.dir,"/mean_gap_filled.txt"),
       std.gap.filled.output.path=paste0(tmp.dir,"/std_gap_filled.txt"),
       rmse.plot.test.data=T, verbose=F)
imputed <- read.table(file="mean_gap_filled.txt", header=T, dec=".", sep="\t")

i=1
for(i in 1:6){
       imputed[,i] <- imputed[,i]*trans_par[i,4]
       imputed[,i] <- imputed[,i]+trans_par[i,3]
       imputed[,i] <- 10^imputed[,i]
}

imputed$T1[which(imputed$T1>max(trait.raw$T1, na.rm=T))]
imputed$T1[which(imputed$T1<min(trait.raw$T1, na.rm=T))]
#... and so on

#determining the number of iterations
trait.info01 <- !(is.na(trait.info))
count <- apply(trait.info01, 1, sum)
```

```
count02 <- count
count02[count02==1]<-0
N.iter <- sum(count02) #the loop for will have N.iter cicles


comp <- data.frame(spp=rep(NA, N.iter), g1=NA, g2=NA, g3=NA, empiric=NA, imputed=NA,
orig_emp=NA, std=NA, trait=NA)


g <- c(seq(from=100, to=sum(count02), by=100), sum(count02))


k=1
i=1
for(i in 1:nrow(trait.info)){
       if(count[i]==1){next}

       trs <- which(!(is.na(trait.info[i,])))
       for(j in 1:length(trs)){
              comp$g0[k] <- as.character(hierarchy.info$g0[i])
              comp$g1[k] <- as.character(hierarchy.info$g1[i])
              comp$g2[k] <- as.character(hierarchy.info$g2[i])
              comp$g3[k] <- as.character(hierarchy.info$g3[i])
              comp$trait[k] <- colnames(trait.info)[trs[j]]

              comp$empiric[k] <- trait.info[i, trs[j]]
              comp$orig_emp[k] <- trait.raw[i, trs[j]]

              #I remove this specific observation
              trait.info02 <- trait.info
              trait.info02[i, trs[j]] <- NA

              #imputation
              GapFilling(trait.info02, hierarchy.info,
                     prediction.level = 4,
                     used.num.hierarchy.levels = 3,
                     mean.gap.filled.output.path = paste0(tmp.dir,"/mean_gap_filled.txt"),
                     std.gap.filled.output.path=paste0(tmp.dir,"/std_gap_filled.txt"),
```

```
                rmse.plot.test.data = FALSE, verbose=F)

        imputed <- read.table(file="mean_gap_filled.txt", header=T, dec=".", sep="\t")
        std <-  read.table(file="std_gap_filled.txt", header=T, dec=".", sep="\t")

        #I save the imputed data
        comp$imputed[k] <- imputed[i, trs[j]]
        comp$std[k] <- std[i, trs[j]]
        if(k %in% g){
                save.image("environment.RData")
                write.csv(comp, "comp.csv")
        }

        print(c("k=",k, ", i=", i))
        k=k+1
    }
}

comp$empRaw <- comp$impRaw <- NA

i=1
for(i in 1:nrow(comp)){
    x <- comp[i,c("empiric", "imputed")]
    j <- NA
    if(comp$trait[i]=="T1"){j=1}
    if(comp$trait[i]=="T2"){j=2}
    if(comp$trait[i]=="T3"){j=3}
    if(comp$trait[i]=="T4"){j=4}
    if(comp$trait[i]=="T5"){j=5}
    if(comp$trait[i]=="T6"){j=6}
    DE <- trans_par[j,4]
    media <- trans_par[j,3]
    min_x <- trans_par[j,2]
    x <- x*DE
    x <- x+media
```

```
       x <- 10^x
       comp[i,c("empRaw", "impRaw")] <- x
print(i)
}
sum(is.na(comp$impRaw)) #0 NAs

#Check if back-transformation is working ok.
plot(comp$empRaw, comp$orig_emp)
       abline(0,1)

#I check if there are imputed values out of the empirical range, and I bound the outliers
#H
maxT1 <- max(comp$orig_emp[comp$trait=="T1"])
minT1 <- min(comp$orig_emp[comp$trait=="T1"])
sum(comp$impRaw[comp$trait=="T1"]>maxT1)
sum(comp$impRaw[comp$trait=="T1"]<minT1)
#... and so on

comp$impRaw[comp$trait=="T1" & comp$impRaw > maxT1] <- maxT1
comp$impRaw[comp$trait=="T1" & comp$impRaw < minT1] <- minT1
#... and so on



#Is the SD reported for every single prediction related to the relative error of each
observation?
comp$error.p <- 100*(comp$impRaw - comp$orig_emp)/comp$orig_emp
plot(abs(comp$error.p)~comp$std, log="xy")



write.csv(comp, "comp.csv")

compT1 <- comp[comp$trait=="T1",]
#... and so on
```

```
Q <- function(obs, imp){
                    complete <- which(!is.na(obs) & !is.na(imp))
                    obs <- obs[complete]
                    imp <- imp[complete]
                    TSS <- sum((obs-mean(obs))^2)
                    RSS <- sum((imp-obs)^2)
                    Ri2 <- 1-(RSS/TSS)
                    Ri2
             }

zRMSE <- function(obs, imp){
                    complete <- which(!is.na(obs) & !is.na(imp))
                    obs <- obs[complete]
                    imp <- imp[complete]
                    rmse <- sqrt((1/length(complete))*sum((obs-imp)^2))
                    rmse
             }

##T1
T1imp_emp01 <- lm(log10(orig_emp)~log10(impRaw), data=compT1)
qqnorm(resid(T1imp_emp01))
       qqline(resid(T1imp_emp01)) #OK

#accuracy
summary(T1imp_emp01) #to know the intercept and slope of the acctual trait value as a
funtion of the imputed value
confint(T1imp_emp01)

#precision
Q(log10(compT1$orig_emp), log10(compT1$impRaw))
zRMSE(compT1$empiric, compT1$imputed)

##T2
#... and so on
```