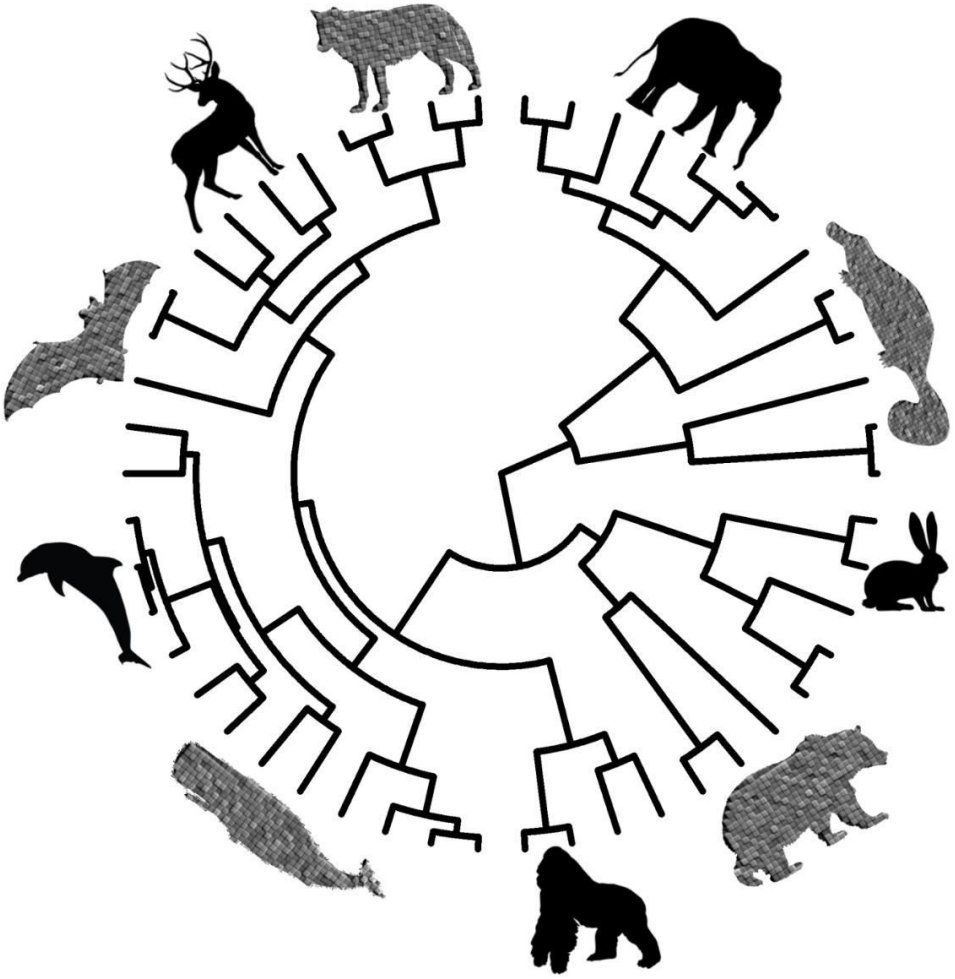
Universidade Federal de Goiás
Instituto de Ciências Biológicas
Programa de Pós-graduação em Ecologia e Evolução

IMPUTAÇÃO FILOGENÉTICA:

UMA PERSPECTIVA MACROECOLÓGICA



LUCAS LACERDA CALDAS ZANINI JARDIM

Goiânia / 2018

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[ ] Dissertação** **[X] Tese**

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Lucas Lacerda Caldas Zanini Jardim

Título do trabalho: Imputação filogenética: uma perspectiva macroecológica

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.

Lucas Lacerda Caldas Zanini Jardim

Assinatura do(a) autor(a)

Ciente e de acordo:

Assinatura do(a) orientador(a)

Data: 12 / 10 / 2018

1

Universidade Federal de Goiás
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Ecologia e Evolução

Lucas Lacerda Caldas Zanini Jardim

IMPUTAÇÃO FILOGENÉTICA: UMA PERSPECTIVA MACROECOLÓGICA

Orientador: José Alexandre Felizola Diniz Filho

Co-orientador: Crisóforo Fabrício Villalobos Camacho

Tese apresentada à Universidade Federal de Goiás como parte das exigências do Programa de Pós-Graduação em Ecologia e Evolução para a obtenção do título de *Doutor em Ecologia e Evolução*.

Goiânia

Abril / 2018

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Jardim, Lucas Lacerda Caldas Zanini
Imputação filogenética: uma perspectiva macroecológica [manuscrito] / Lucas Lacerda Caldas Zanini Jardim. - 2018.
128 f.: il.

Orientador: Prof. Dr. José Alexandre Felizola Diniz-Filho; co orientador Dr. Crisóforo Fabrício Villalobos Camacho .
Tese (Doutorado) - Universidade Federal de Goiás, , Programa de Pós-Graduação em Ecologia e Evolução, Goiânia, 2018.
Bibliografia.

1. Imputação múltipla. 2. imputação filogenética. 3. macroecologia. 4. regra de Bergmann. 5. Homo floresiensis. I. Diniz-Filho, José Alexandre Felizola , orient. II. Título.

CDU 574

SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DE GOIÁS - UFG
INSTITUTO DE CIÊNCIAS BIOLÓGICAS - ICB
PROGRAMA DE PÓS-GRADUAÇÃO EM ECOLOGIA E EVOLUÇÃO - PPGEE

ATA DA SESSÃO PÚBLICA DE DEFESA DE TESE Nº 67

Aos vinte e sete dias do mês de abril de 2018 (27/04/2018), às nove horas (09h), no Auditório do ICB V, UFG, reuniram-se os componentes da banca examinadora: **Prof. Dr. José Alexandre Felizola Diniz Filho, ICB-UFG; Prof. Dr. João Carlos Nabout, UEG-Anápolis; Prof. Dr. Matheus de Souza Lima Ribeiro, ICB-UFG; Prof. Dr. Daniel de Paiva Silva, IFG-URUTAÍ; , Prof. Dr. Tiago Bosisio Quental, USP**; para, em sessão pública presidida pelo (a) primeiro(a) examinador(a) citado(a), procederem à avaliação da defesa de tese intitulada: **"Imputação filogenética: uma perspectiva macroecológica"**, em nível de doutorado, área de concentração em Ecologia e Evolução, de autoria de **Lucas Lacerda Caldas Zanini Jardim**, discente do Programa de Pós-Graduação em Ecologia e Evolução da Universidade Federal de Goiás. A sessão foi aberta pelo(a) presidente(a), que fez a apresentação formal dos membros da banca. A palavra, a seguir, foi concedida a(o) autor(a) da tese que, em cerca de 30 minutos, procedeu à apresentação de seu trabalho. Terminada a apresentação, cada membro da banca arguiu a(o) examinada(o), tendo-se adotado o sistema de diálogo sequencial. Terminada a fase de arguição, procedeu-se à avaliação da tese. Tendo-se em vista o que consta na Resolução nº 1127 de dezembro de 2012 do Conselho de Ensino, Pesquisa, Extensão e Cultura (CEPEC), que regulamenta o Programa de Pós-Graduação em Ecologia e Evolução, a tese foi aprovada, considerando-se integralmente cumprido este requisito para fins de obtenção do título de Doutor(a) em Ecologia e Evolução pela Universidade Federal de Goiás. A conclusão do curso dar-se-á quando da entrega da versão definitiva da tese na secretaria do programa, com as devidas correções sugeridas pela banca examinadora, no prazo de trinta dias a contar da data da defesa. Cumpridas as formalidades de pauta, às 13 h e

00 min., encerrou-se a sessão de defesa e, para constar, eu, Suely Ana Ribeiro, secretária executiva da Universidade Federal de Goiás - UFG, lavrei a presente ata que, após lida e aprovada, será assinada pelos membros da banca examinadora em três vias de igual teor.

Prof. Dr. José Alexandre Felizola Diniz Filho
Presidente da banca
ICB-UFG

Prof. Dr. João Carlos Nabout
UEG-Anápolis

Prof. Dr. Matheus de Souza Lima Ribeiro
ICB-UFG

Prof. Dr. Daniel de Paiva Silva
IFG-URUTAÍ

Prof. Dr. Tiago Bosisio Quental
USP

AGRADECIMENTOS

Ao longo dos 4 anos de doutorado acumulei um enorme débito de gratidão com muitas pessoas que me aconselharam, apoiaram, corrigiram e que foram referências de como deve ser um profissional na carreira científica. O meu primeiro agradecimento é à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que graças à concessão da bolsa de doutorado, pude dedicar-me exclusivamente ao desenvolvimento dessa tese.

Aos meus amigos e orientadores **José Alexandre Felizola Diniz Filho** e **Fabrício Villalobos** que sempre estiveram presentes e souberam me aconselhar e guiar ao longo desses anos. Hoje, com certeza, sou uma pessoa melhor graças a vocês.

Agradeço ao **Daniel Brito** por ter me orientado durante o meu mestrado, **Natan Maciel** por ter avaliado os meus relatórios de acompanhamento e **Thiago Rangel** e **Luiz Bini** por terem participado da minha qualificação. Aos outros professores do PPG, agradeço pelos ensinamentos nas disciplinas, palestras e estágios docentes.

Tenho também que agradecer a todos os amigos que fiz durante a minha vida acadêmica. Começando pelos amigos de Alfenas, **Ed**, **Arildo**, **Bruno**, **Renan, Cássio, Márcio, Rodolpho Rodrigues, Ricardo Marcelino, Mariana Raniero, Mainara, Marcela, Kelson**, **Marcel, Thais** e tantos outros que a minha memória e o prazo curto para escrever me impediram de lembrar. Devo também agradecer ao **Rogério Grasseto** e **Érica Hasui** que me orientaram e introduziram na vida científica.

Em Goiânia fiz muitos amigos como **Franciele Parreira**, **Artur Bispo** e **Diogo Provete**, que me ajudaram no início. Com certeza, não posso esquecer de **Bruno Vilela**, **Marcos Vieira**, **Edmar Almeida**, **Fernando Landa**, **Frederico Faleiro**, **Renan Costa**, **Jaques Zanon**, **Ivy Gobeti**, **Luciano Sgarbi**, **Lorena Simon**, **Raísa Vieira**, **Danilo Fortunato**, **Bruno Ribeiro**, **José Hidasi Neto**, **Davi Alves**, **Jesus Pinto-Ledezma**, **Paola Nobre**, **Lucas Gontijo**, **Priscila Cabral**, **Cléber Ten Caten**, **Dainel Paiva**, **Alice Francener**, **Kelly Souza**, **Flávia Machado**, **Elisa Barreto**, **Lucas Camargo**, **Rodolfo Cabral**, **Ludmila Rattis**, **Carol Caiado**, **Fernando Roa**, **Bruno Barreto**, **Vinícius Guerra**, **Renato Dala Corte**, **Daniel Plazas-Jimenez**, **Anderson Medina**, **Karlo Guidoni** e **Rômulo Pinto**. Tenho plena certeza que outras pessoas merecem também

meus agradecimentos e não foram incluídas na lista acima, peço que perdoem a minha memória.

Por fim, agradeço aos meus pais, **Ricardo Zanini Jardim** e **Suzana Maria Lacerda Caldas Jardim** pelo incentivo e por terem sempre se sacrificado pela garantia da minha educação. Aos meus irmãos, **Bernardo** e **Bárbara**, pela amizade e parceria. Meus **primos**, **avós** e **tios**, que sempre participaram do meu crescimento. Agradeço também ao **Júlio Abreu Jr.**, **Marco Antônio**, **Jady Fernandes**, **Vania Ferrari**, **Antônio Silva**, **Magali** e **Júlio Abreu**. Finalmente, devo muito à minha esposa **Tatianne Piza Ferrari Abreu Jardim**, que sempre tornou os momentos difíceis mais suaves, pela ajuda intelectual e editorial e por ter aceitado construir uma vida comigo.

SUMÁRIO

RESUMO GERAL **7**

GENERAL ABSTRACT **8**

INTRODUÇÃO GERAL **9**

REFERÊNCIAS 11

CAPÍTULO 1: ***Challenging the Raunkiaeran shortfall and the consequences of using imputed databases*** **16**

ABSTRACT 16

INTRODUCTION 17

METHODS 21

Phylogeny simulation 21

Trait simulation 21

Missing data scenarios 23

Imputation methods 24

Estimating phylogenetic signal 25

Imputation effects on phylogenetic signal and descriptive statistics 25

Imputation error 26

Overall analyses 26

RESULTS 27

DISCUSSION 33

CONCLUDING REMARKS 41

ACKNOWLEDGEMENTS 42

REFERENCES 42

APPENDIX 1 49

APPENDIX 2 57

CAPÍTULO 2: ***Phylogenetic imputation and brain-body size evolution in primates, with special reference to Homo floresiensis*** **62**

ABSTRACT 62

INTRODUCTION 63

METHODS 63
Insularity definition, body and brain size 63
Phylogenetic hypothesis 68
Body and brain size evolutionary models 68
Model adequacy 71
Predicting Homo floresiensis body and brain size 73
RESULTS 74
Body size evolutionary model 74
Brain size evolutionary model 75
DISCUSSION 79
Body size evolution and Island Rule in primates 79
Brain size evolution 86
CONCLUSION 88
ACKNOWLEDGEMENTS 89
LITERATURE CITED 89

CAPÍTULO 3: *Cross-species evaluation of Bergmann's rule in mammals: looking at biodiversity knowledge shortfall implications* 97

ABSTRACT 97
INTRODUCTION 98
MATERIAL AND METHODS 101
Database and phylogenies 101
Multiple imputation 102
Statistical analyses 105
RESULTS 105
DISCUSSION 106
CONCLUSION 113
ACKNOWLEDGEMENTS 114
REFERENCES 114
SUPPLEMENTARY MATERIAL 126
CONCLUSÂO GERAL 128

Resumo geral

A macroecologia estuda padrões ecológicos em grandes escalas geográficas e temporais, em busca de quais processos moldam esses padrões. Nessas escalas de estudo, há raramente informações completas sobre as centenas ou até milhares de espécies estudadas. Essa ausência de informações tem o potencial de enviesar as conclusões dos estudos sobre padrões e processos macroecológicos. Nessa tese, nós avaliamos métodos de imputação filogenética, a sua aplicação e consequências em estudos macroecológicos. Para avaliar potenciais vieses do uso de banco de dados imputados, no primeiro capítulo, nós aplicamos diferentes métodos utilizados para tratar dados faltantes, sob diferentes cenários de evolução dos atributos das espécies, porcentagem e padrão dos dados faltantes. Nós encontramos que a forma de tratar o dado faltante pode ser dependente dos objetivos e dos dados de cada estudo e, portanto, nós sugerimos cautela ao utilizarmos bancos de dados imputados. No segundo capítulo, nós testamos o efeito da regra de ilha na evolução da massa corpórea e do volume cerebral de primatas. A partir dos melhores modelos evolutivos ajustados a esses atributos, nós imputamos a massa corpórea e volume cerebral de *Homo floresiensis*. Nós concluímos que primatas não seguem regra de ilha e que apesar de nossos modelos superestimarem, em média, o tamanho do corpo e cérebro de *Homo floresiensis*, a sua evolução não se desvia do esperado pela evolução de primatas. Por fim, no terceiro capítulo testamos a regra de Bergmann em mamíferos, utilizando métodos de imputação múltipla e avaliamos as consequências de desconsiderar os dados faltantes na detecção da regra. Nós encontramos que testar a regra sem considerar os dados faltantes pode inverter o efeito da temperatura na massa do corpo, mas esse viés não tornou o efeito estatisticamente significante. Portanto, concluímos que mamíferos não seguem a regra de Bergmann, quando toda a classe é avaliada. Por fim, essa tese discutiu vantagens, desvantagens e futuras linhas de pesquisa para tornar a imputação filogenética uma ferramenta mais robusta para tratarmos dados faltantes em macroecologia.

Palavras-chave: Imputação múltipla, imputação filogenética, macroecologia, dados faltantes, lacuna de conhecimento, regra de Bergmann, regra de ilha, *Homo floresiensis*

General abstract

Macroecology studies ecological pattern at large geographical and temporal scales. At these scales, information about hundreds or even thousands of studied species. This lack of information may potentially bias studies' conclusions related with macroecological processes and patterns. In this thesis, we evaluated phylogenetic imputation methods, their uses and effects in macroecological studies. The first chapter evaluated different methods used to deal with missing data, taking into account different scenarios of species trait evolution, as well as percentage and pattern of missing data. We found that dealing with missing data relies on the specific goals and data of the study. Therefore, we suggested caution while using imputed database. In the second chapter, we tested the island rule effect in body mass and brain volume of primates. To do so, we fitted evolutionary models to those traits and then imputed the body mass and brain volume for *Homo floresiensis*. We concluded that primates do not follow the island rule and even though our models overestimated, on average, brain and body size of *Homo floresiensis*, its evolution did not deviate from primates' evolutionary expectation. Lastly, in the third chapter, we tested existence of Bergmann's rule in mammals using multiple imputation methods, in addition to considering the consequences of ignoring missing data while testing the rule. We found that ignoring missing data can invert (eg. changing from positive to negative effect) the effect of temperature on body mass, but this bias did not turn the effect statistically significant. Therefore, we concluded that mammals do not follow Bergmann's rule, when evaluated at the class taxonomic level. Finally, this thesis discussed pros, cons and future research avenues in order to make phylogenetic imputation a more robust tool to deal with missing data in macroecology.

Keywords: multiple imputation, phylogenetic imputation, macroecology, missing data, biodiversity knowledge shortfall, Bergmann's rule, Island rule, *Homo floresiensis*, mammals

Introdução geral

A macroecologia, como um programa de pesquisa, busca entender quais processos moldam os padrões bióticos que emergem em grandes escalas espaciais e temporais, bem como as suas relações com variáveis abióticas (Brown & Maurer, 1989; Brown, 1995; Smith *et al.*, 2008). Dado esse foco em "grandes escalas", a macroecologia frequentemente necessita de informações sobre centenas ou milhares de espécies, tais como onde elas ocorrem, quais as suas características morfológicas, fisiologia, nichos e abundâncias populacionais (Brown, 1995; Gaston & Blackburn, 2000; Diniz-Filho *et al.*, 2009; Fritz *et al.*, 2009; Clarke *et al.*, 2010; Jetz *et al.*, 2012; Pyron & Wiens, 2013; Mazel *et al.*, 2014; Oliveira *et al.*, 2016a; Villalobos *et al.*, 2016). No entanto, o esforço de pesquisa não é igualmente distribuído, ou aleatório, entre as espécies (Reddy and Dávalos 2003; Cardoso et al. 2011; Costello et al. 2013; Oliveira et al. 2016). Consequentemente, a distribuição dessas informações é normalmente enviesada para algumas espécies e quanto mais aumentamos o número de espécies ou variáveis a serem estudadas, maior é a presença de dados faltantes (Gonzalez-Suarez *et al.*, 2012).

À ausência de informações sobre as espécies tem sido atribuído o termo "lacuna de conhecimento" (Cardoso *et al.*, 2011; Hortal *et al.*, 2015) e há várias dessas lacunas descritas na literatura, comumente nomeadas em homenagem aos pesquisadores ilustres de uma determinada área de pesquisa. Assim, a ausência de conhecimento sobre as ocorrências das espécies foi nomeada como "Lacuna Wallaceana", o desconhecimento em relação à existência de uma espécies é a "Lacuna Linneana" e ainda temos as lacunas "Darwiniana", "Prestoniana", "Raunkiaeriana", "Hutchinsoniana" e "Eltoniana" (Hortal *et al.*, 2015). Essas lacunas representam, respectivamente, a ignorância em relação a história evolutiva das espécies, seus dados populacionais, atributos funcionais, nicho climático e interações bióticas (Hortal *et al.*, 2015).

Todas essas lacunas possuem déficits de informação classificados em três categorias: (1) há conhecimento sobre a existência da informação, mas os dados ainda não foram coletados, (2) os dados já foram coletados, mas são de difícil acesso e (3) não há conhecimento sobre nem mesmo a existência da informação ou fenômeno (Jackson, 2012; Hortal *et al.*, 2015). Assim, as únicas formas de informações faltantes que podem ser tratadas são aquelas cuja existência é conhecida, ou seja, caso (1) e, em algumas situações, caso (2). Portanto, uma vez que há conhecimento sobre a existência das informações, elas podem ser tratadas como dados faltantes.

À medida que as lacunas de conhecimento são assumidas como dados faltantes, todo o arcabouço estatístico desenvolvidos desde os anos 70 para a análise de dados faltantes (Rubin, 1976; Little & Rubin, 2002; Enders, 2010; Molenberghs *et al.*, eds, 2014) torna-se aplicável. Para isso, assumimos que a ausência de dados é uma variável aleatória, ou em outras palavras, durante o processo de coleta de dados há uma probabilidade da informação ser coletada ou passar desapercebida. Essa probabilidade pode ser uniformemente distribuída entre os dados a serem coletados, ou determinada por alguma outra variável conhecida ou até mesmo por uma causa desconhecida. Ao modelarmos corretamente a ausência de dados, as inferências tornam-se menos enviesadas (Rubin, 1976).

Sob esse arcabouço, diferentes abordagens estatísticas foram propostas, como métodos bayesianos e de máxima verossimilhança (Little & Rubin, 2002; Enders, 2010; Molenberghs *et al.*, eds, 2014), mas, o mais flexível deles é a Imputação Múltipla (van Buuren, 2012; Nakagawa, 2015; Murray, 2018), uma vez que a geração dos valores imputados é separada das análises subsequentes (Rubin, 1996). Na Imputação Múltipla, simula-se uma distribuição de possíveis valores para os dados ausentes, garantindo as relações existentes entre as variáveis (Rubin, 1996; Schafer & Graham, 2002; van Buuren, 2012), bem como a reprodução da sua distribuição de frequência, e assim ao analisarmos essa distribuição de valores geramos inferências não enviesadas (Rubin, 1987, 1996).

Ao longo da história, a macroecologia não tem incluído frequentemente os métodos desenvolvidos para tratar dados faltantes nas suas metodologias analíticas, assim como a ecologia e evolução como um todo (Nakagawa & Freckleton, 2008; Nakagawa, 2015). Recentemente, Swenson (2014) propôs a "imputação filogenética", uma ferramenta que utiliza métodos comparativos filogenéticos (Diniz-Filho *et al.*, 1998; Garland, Jr., & Ives, 2000) para o preenchimento de informações faltantes em bancos de dados macroecológicos. Entretanto, a imputação filogenética tem tido como principal objetivo a capacidade de predizer acuradamente os valores faltantes (Guénard *et al.*, 2013; Penone *et al.*, 2014; Swenson, 2014; Diniz-Filho *et al.*, 2015; Molina-Venegas *et al.*, 2018). Por outro lado, a imputação múltipla tem como objetivo recuperar as estruturas do dado de interesse do dado, como exemplo as relações entre as variáveis, suas distribuições de frequência (Rubin, 1996; Schafer & Graham, 2002; van Buuren, 2012) ou padrões macroecológicos. Assim, com o recente interesse no uso de imputações nos estudos ecológicos e evolutivos (Penone *et al.*, 2014; Taugourdeau *et al.*, 2014; Schrodt *et al.*,

2015; Goolsby *et al.*, 2016; Legendre *et al.*, 2016; Molina-Venegas *et al.*, 2018; Swenson *et al.*, 2017), o nosso objetivo nessa tese foi avaliar e aplicar o uso de imputações filogenéticas na inferência de padrões macroecológicos.

Portanto, no primeiro capítulo, nós simulamos diferentes cenários de dados faltantes, com diferentes graus de estruturação filogenética. Em seguida, avaliamos o impacto de diferentes técnicas de tratamento do dado faltante, incluindo imputações múltiplas e filogenéticas, na inferência de diferentes propriedades dos dados faltantes, como sua média, variância, relação com outra variável e sinal filogenético. Concluímos que o uso de dados imputados deve ser específico para os objetivos e dos dados disponíveis para o estudo. Portanto, o uso indiscriminado de bancos de dados imputados sem o conhecimento do processo de imputação e seus impactos no objetivo do estudo podem causar conclusões enviesadas.

No segundo capítulo, nós aplicamos a imputação filogenética para testar a hipótese sobre o efeito de ilha no tamanho do corpo e cérebro de *Homo floresiensis*. Para isso, nós ajustamos diferentes modelos evolutivos sobre como tamanho do corpo e do cérebro evoluíram na ordem dos primatas e assim imputamos esses atributos em *Homo floresiensis*, testando se esse hominídeo desviou-se ou não do que seria esperado pela evolução dos primatas. Nós encontramos que primatas não seguem a regra de ilha e que o tamanho do corpo e cérebro de *Homo floresiensis* estão dentro do esperado pela evolução dos primatas.

No terceiro capítulo, nós avaliamos como desconsiderar dados faltantes podem enviesar as inferências macroecológicas, e para isso, usamos como estudo de caso o teste da regra de Bergmann em mamíferos. Para a avaliação do efeito do dado faltante, nós contrastamos as análises que deletam espécies sem dados e com aquelas que realizam imputações múltiplas filogenéticas. Nós concluímos que a regra de Bergmann não se aplica aos mamíferos, quando avaliado ao nível de classe. Além disso, encontramos que a desconsideração dos dados faltantes pode enviesar o efeito da temperatura na massa corpórea de mamíferos, mas esse viés não mudou nossas conclusões sobre a ausência da regra.

Referências

Brown, J.H. 1995. *Macroecology*, 1st edition. University of Chicago Press, Chicago.

Brown, J.H. & Maurer, B.A. 1989. Macroecology: the division of food and space among species on continents. *Science.* **243**: 1145–1150.

Cardoso, P., Erwin, T.L., Borges, P. a. V. & New, T.R. 2011. The seven impediments in invertebrate conservation and how to overcome them. *Biol. Conserv.* **144**: 2647–2655.

Clarke, A., Rothery, P. & Isaac, N.J.B. 2010. Scaling of basal metabolic rate with body mass and temperature in mammals. *J. Anim. Ecol.* **79**: 610–619.

Costello, M.J., May, R.M. & Stork, N.E. 2013. Can we name Earth's species before they go extinct? *Science.* **339**: 413–416.

Diniz-Filho, J.A.F., Rodríguez, M.Á., Bini, L.M., Olalla-Tarraga, M.Á., Cardillo, M., Nabout, J.C., *et al.* 2009. Climate history, human impacts and global body size of Carnivora (Mammalia: Eutheria) at multiple evolutionary scales. *J. Biogeogr.* **36**: 2222–2236.

Diniz-Filho, J.A.F., Sant'Ana, C.E.R. & Bini, L.M. 1998. An eigenvector method for estimating phylogenetic inertia. *Evolution.* **52**: 1247–1262.

Diniz-Filho, J.A.F., Villalobos, F. & Bini, L.M. 2015. The best of both worlds : Phylogenetic eigenvector regression and mapping. *Genet. Mol. Biol.* **38**: 396–400.

Enders, C.K. 2010. *Applied missing data analysis*, 1st edition. Guilford Press.

Fritz, S. a, Bininda-Emonds, O.R.P. & Purvis, A. 2009. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecol. Lett.* **12**: 538–49.

Garland, Jr., T. & Ives, A.R. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat.* **155**: 346–364.

Gaston, K.J. & Blackburn, T.M. 2000. *Pattern and process in macroecology*, 1ª edition. Wiley-Blackwell.

Gonzalez-Suarez, M., Lucas, P.M. & Revilla, E. 2012. Biases in comparative analyses of extinction risk: mind the gap. *J. Anim. Ecol.* **81**: 1211–22.

Goolsby, E.W., Bruggeman, J. & Ané, C. 2016. Rphylopars : fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods Ecol. Evol.* **8:** 22-27

Guénard, G., Legendre, P. & Peres-Neto, P. 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. *Methods Ecol. Evol.* **4**: 1120–1131.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **46**: 523–549.

Jackson, S.T. 2012. Representation of flora and vegetation in Quaternary fossil assemblages: known and unknown knowns and unknowns. *Quat. Sci. Rev.* **49**: 1–15.

Jetz, W.T., Joy, G.H., Hartmann, J.B., Mooers, K., Thomas, G.H., Joy, J.B., *et al.* 2012. The global diversity of birds in space and time. *Nature* **491**: 444–8.

Legendre, L.J., Guénard, G., Botha-brink, J. & Cubo, J. 2016. Palaeohistological evidence for ancestral high metabolic rate in archosaurs. *Soc. Syst. Biol.* **0**: 1–26.

Little, R.J.A. & Rubin, D.B. 2002. *Statistical analysis with missing data*, 2nd editio. John Wiley & Sons.

Mazel, F., Guilhaumon, F., Mouquet, N., Devictor, V., Gravel, D., Renaud, J., *et al.* 2014. Multifaceted diversity-area relationships reveal global hotspots of mammalian species, trait and lineage diversity. *Glob. Ecol. Biogeogr.* **23**: 836–847.

Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A. & Verbeke, G. (eds). 2014. *Handbook of Missing Data Methodology*, 1º ed. CRC Press.

Molina-Venegas, R., Moreno-Saiz, J.C., Castro Parga, I., Davies, T.J., Peres-Neto, P.R. & Rodríguez, M.A. 2018. Assessing among-lineage variability in phylogenetic imputation of functional trait datasets. *Ecography.* 1–10.

Murray, J.S. 2018. Multiple imputation : A review of practical and theoretical findings. arXiv:1801.04058.

Nakagawa, S. 2015. Missing data: mechanisms, methods, and messages. In: *Ecological statistics: contemporary theory and application* (G. A. Fox et al., eds), pp. 81–105.

Oxford University Press.

Nakagawa, S. & Freckleton, R.P. 2008. Missing inaction: the dangers of ignoring missing data. *Trends Ecol. Evol.* **23**: 592–596.

Oliveira, B.F., Machac, A., Costa, G.C., Brooks, T.M., Davidson, A.D., Rondinini, C., *et al.* 2016a. Species and functional diversity accumulate differently in mammals. *Glob. Ecol. Biogeogr.* **25**: 1119–1130.

Oliveira, U., Paglia, A.P., Brescovit, A.D., de Carvalho, C.J.B., Silva, D.P., Rezende, D.T., *et al.* 2016b. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Divers. Distrib.* 1–13.

Penone, C., Davidson, A.D., Shoemaker, K.T., Marco, M. Di, Rondinini, C., Brooks, T.M., *et al.* 2014. Imputation of missing data in life-history traits datasets: which approach performs the best? *Methods Ecol. Evol.* **5**: 961–970.

Pyron, R.A. & Wiens, J.J. 2013. Large-scale phylogenetic analyses reveal the causes of high tropical amphibian diversity. *Proc. Biol. Sci.* **280**: 20131622.

Reddy, S. & Dávalos, L.M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *J. Biogeogr.* **30**: 1719–1727.

Rubin, D.. 1976. Inference and missing data. *Biometrika* **63**: 581–592.

Rubin, D.B. 1996. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **91**: 473.

Rubin, D.B. 1987. *Multiple imputation for nonresponse in surveys*, 1st ed. John Wiley & Sons.

Schafer, J.L. & Graham, J.W. 2002. Missing data: our view of the state of the art. *Psychol. Methods* **7**: 147–177.

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., *et al.* 2015. BHPMF - a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Glob. Ecol. Biogeogr.* **24**: 1510–1521.

Smith, F.A., Lyons, S.K., Ernest, S.K.M. & Brown, J.H. 2008. Macroecology : more than the division of food and space among species on continents. *Prog. Phys. Geogr.* **32**: 115–138.

Swenson, N.G. 2014. Phylogenetic imputation of plant functional trait databases.

Ecography. **37**: 105–110.

Swenson, N.G., Weiser, M.D., Mao, L., Araújo, M.B., Diniz-Filho, J.A.F., Kollmann, J., *et al.* 2017. Phylogeny and the prediction of tree functional diversity across novel continental settings. *Glob. Ecol. Biogeogr.* **26**: 553–562.

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O. & Amiaud, B. 2014. Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data. *Ecol. Evol.* **4**: 944–958.

van Buuren, S. 2012. *Flexible imputation of missing data*, 1st edition. Chapman and Hall/CRC.

Villalobos, F., Olalla-Tárraga, M.Á., Cianciaruso, M. V., Rangel, T.F. & Diniz-Filho, J.A.F. 2016. Global patterns of mammalian co-occurrence: phylogenetic and body size structure within species ranges. *J. Biogeogr.* 1–11.

Challenging the Raunkiaeran shortfall and the consequences of using imputed databases

Lucas Jardim[1]*, Luis Mauricio Bini[1], José Alexandre Felizola Diniz-Filho[1], Fabrício Villalobos[1,2]

[1]Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Goiás, Brasil;

[2]Red de Biología Evolutiva, Instituto de Ecología, A.C., Carretera antigua a Coatepec 351, El Haya, 91070 Xalapa, Veracruz, Mexico

*Corresponding author: Lucas Jardim, Departamento de Ecologia, ICB, Universidade Federal de Goiás, 74690-900, Goiânia, Goiás, Brasil. E-mail: lucas.ljardim9@gmail.com.

Running title: *Phylogenetic signal and trait imputed databases*

Abstract

Given the prevalence of missing data on species' traits – Raunkiaeran shortfall-, several methods have been proposed to fill sparse databases. Analyses based on these imputed databases can introduce several biases. Here, we evaluated potential biases in descriptive statistics, regression parameters, and phylogenetic signal estimated from imputed databases under different missing and imputing scenarios. We found that percentage of missing data, missing mechanisms, Ornstein-Uhlenbeck strength and handling methods were important in determining errors of estimates. We also found that imputation errors are not linearly related to estimate errors. Although without biases, adding phylogenetic information provides better estimates of evaluated parameters. We advise researchers to share both their raw and imputed data and users to consider the pattern of missing data to

find methods that overcome this problem before running their analyses. In addition, new developments of phylogenetic methods should consider imputation uncertainty, phylogenetic autocorrelation and phylogenetic structure of the original data.

Key-words: bias, Multiple Imputation, trait databases, Phylogenetic Eigenvector Maps, phylogenetic signal, Phylogenetic Comparative Methods.

Introduction

Missing data are a ubiquitous feature of real-world datasets (Nakagawa & Freckleton 2008). Lack of information may limit the application of statistical analysis and can lead to biased estimates and conclusions on the phenomena of interest. In 1976, Donald B. Rubin proposed a missing-data theory to allow analysis of incomplete datasets (Rubin, 1976), explaining how unbiased parameters could be estimated with missing data by considering the mechanisms causing missing data. These mechanisms were classified into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). They mean, respectively, that missing values are equally probable across a dataset, probability of missing data is correlated with other variables rather than to the variable with missing data (target variable), and probability of missing data is itself correlated to the target variable and dependent on the missing data (Rubin, 1976; Nakagawa & Freckleton, 2008; Enders, 2010; van Buuren, 2012) (Fig.1).

When dealing with missing data, the above mechanisms need to be taken into account before analysis (Rubin 1976). This is because different methods that handle missing data assume different mechanisms, so using them indiscriminately may bias

Figure 1. Correlation structure among variables in each missing data mechanism. Circles represent model components and their intersection represents correlation among them.

parameter estimates (Rubin, 1976; Enders, 2010; van Buuren, 2012). Multiple Imputation and Full Information Maximum Likelihood methods are currently regarded as the most appropriate methods to handle missing data, because they work under MAR and MCAR scenarios and provide unbiased estimates (Enders, 2010). In contrast, it is very difficult to model missing data under a MNAR scenario. This is so due to the need of considering a model that represents the probability of missing values to occur and because the shape of the probability density function is not known (Enders, 2010; van Buuren, 2012).

Research in ecology and evolutionary biology usually requires data about species and their traits to answer different questions from community assembly and ecogeographical rules to correlated evolution, diversification rates and extinction probability, among others (Purvis et al., 2000; Webb et al., 2002; Gaston et al., 2008; Goldberg et al., 2010; Lukas & Clutton-Brock, 2013; Jetz & Freckleton, 2015). Thus, to facilitate research and make it reproducible and data more accessible (Reichman et al., 2011), ecologists and evolutionary biologists usually create databases that include information on huge amounts of species and their traits (e.g., (Jones et al., 2009; Kattge

et al., 2011; Wilman et al., 2014). However, as databases become larger, the probability of having all the necessary data for all species rapidly decreases. This lack of knowledge about species' traits and their ecological functions was recently defined as the Raunkiaeran shortfall (Hortal et al., 2015) or Eltonian shortfall (Rosado et al., 2015).

Owing to the ubiquity of the Raunkiaeran shortfall, some researchers are interested in filling such gaps in their databases for their own analyses but also to make them available for other researchers (Swenson, 2014; Schrodt et al., 2015). To do so, recent studies suggest the use of phylogenetic information in the imputation process (Guénard et al., 2013; Swenson, 2014; Schrodt et al., 2015). Phylogenetic information is important in imputation because closely related species resemble, on average, each other more than distantly related species. This phenomenon is commonly known as phylogenetic signal (Blomberg et al., 2003). Consequently, knowing the phylogenetic position of species could, in principle, be used to perform a good estimation of missing trait values. However, the relationship between trait divergence and phylogenetic distance may be more complex (due to distinct evolutionary models) than usually assumed (Hansen & Martins, 1996; Münkemüller et al., 2012). For instance, under an Ornstein-Uhlenbeck evolutionary model traits may evolve under selection restrictions where species track a trait optimum, causing phenotypic resemblance even among phylogenetically distant species (Hansen & Martins, 1996). Alternatively, under an Early-burst model traits may show evolutionary rates early in species history and later the rates slow down, resulting in phylogenetically closely related species having different trait values (Blomberg et al., 2003; Harmon et al., 2010). Finally, trait evolution may behave like a drift process (e.g., Brownian motion) where species trait differences are directly correlated with time since divergence (Felsenstein, 1985; Hansen & Martins, 1996; Freckleton et al., 2002). Therefore, imputation methods should explicitly consider or

assume a trait evolutionary model determining the relationship between species resemblance and phylogenetic proximity (Guénard et al., 2013).

Nowadays, large, imputed databases already exist that used taxonomic, ecological or allometric relationships to fill in missing values (Jones et al., 2009; Wilman et al., 2014). This highlights the need to critically evaluate the use of imputed databases given that the reliability of statistical analysis under missing data is dependent on how many values were missing in the original data, what mechanism caused data to be missing and which methods were used in the imputation process (Schafer & Graham, 2002; Enders, 2010; van Buuren, 2012). Moreover, other problems can also arise when testing for phylogenetic signal (Cavender-Bares et al., 2009; Münkemüller et al., 2012). In such cases, if analysis were to be conducted on phylogenetically imputed data, results could be misleading given that missing values would have been already filled based on their phylogenetic structure, thus potentially inflating the level of phylogenetic signal. This potential issue can have important consequences for studies evaluating, for example, niche conservatism, trait lability, community assembly and diversification (Blomberg et al., 2003; Wiens & Graham, 2005; Cavender-Bares et al., 2009; Goldberg et al., 2010).

Considering the current need for complete databases and the use of imputation methods to accomplish this, we evaluate how the estimation of descriptive statistics, regression coefficients and phylogenetic signal can be misled by the percentage of missing data, the particular mechanism of missing data, the model of trait evolution and the choice of methods used to handle missing values. To accommodate these scenarios, we use simulated phylogenies and traits under different combinations of such conditions. In addition, to address imputation accuracy, we evaluated the relationship between error caused by imputation and estimate errors.

Methods

Phylogeny simulation

To evaluate the effect of imputing missing values into sparse databases (i.e. with missing data), we first simulated 100 birth-death phylogenies, with speciation and extinction rates respectively equal to 1 and 0. Each phylogeny had 200 species and were simulated using the function *pbtree* from the R package *phytools* (Revell, 2012). We focused on this phylogeny size because it has been considered appropriate to evaluate power and accuracy of phylogenetic analysis (Davis et al., 2013; Cooper et al., 2016), and it represents a conservative approximation to database size (e.g. several hundreds to thousands of species).

Trait simulation

For each phylogeny, we simulated two traits: a target trait and an auxiliary trait. The first trait represented the one that would be imputed (i.e. missing-value trait), whereas the second trait represented an auxiliary trait that would be used to impute values for the target trait.

The target trait was simulated using the *rTraitCont* function from the *ape* package (Paradis et al., 2004). We modeled this trait under a Ornstein-Uhlenbeck evolutionary process (OU) (Gillespie, 1996), because it allowed us to simulate trait evolution within a continuum from evolutionary drift (i.e. Brownian motion) to weak and strong levels of selection strength on trait evolution (Hansen & Martins, 1996; Hansen, 1997). Thus, we could evaluate the performance of imputation methods under different levels of phylogenetic signal. We fixed the target trait's optimum (Θ) to zero and the trait interspecific variation (σ) equal to one. Also, we simulated different selection strengths

by varying α (selective strength) from 0 to 2, in 0.5 steps (0, 0.5, 1, 1.5 and 2). Such values covered evolutionary scenarios from Brownian motion (OU α = 0) to strong selective strength (OU α = 2).

The auxiliary trait represented a variable used to impute values into the target trait. We simulated auxiliary traits in two ways: (i) correlated with the phylogeny and (ii) correlated with the target trait but uncorrelated with phylogeny. For (i), we simulated the trait following Liam Revell (pers. comm.):

$$x = ry + \sqrt{1 - r^2}\, MVN(0, \sigma^2 \sum) \qquad \text{eqn 1}$$

where *y* is the target trait, *x* the auxiliary trait, and *r* the correlation coefficient between both traits, which was set to *r* = 0.6. We also performed all analyses with *r* = 0.9 to explore the sensibility of our results to the strength of trait correlation, but showed in the main text only results for *r* = 0.6 (see results below). ∑ is the species covariance matrix (Felsenstein, 1985; Revell et al., 2008) and σ² the target trait variation rate calculated as the mean of squared phylogenetic independent contrasts (Freckleton & Jetz, 2009), which was estimated using the *pic* function from *ape (Paradis et al., 2004)*. MVN means Multivariate Normal Distribution and it was simulated using the *fastBM* function from the *phytools* R package (Revell, 2012). This auxiliary trait was later used when simulating the MCAR (Missing Completely at Random) and MAR.PHYLO (Missing at Random correlated with phylogeny) (see below).

For the (ii) scenario, where the auxiliary trait is correlated with the target trait but uncorrelated with phylogeny, the auxiliary trait was simulated using equation 1 with ∑ having off-diagonal entries equal to zero (i.e. no covariance among species) and diagonal entries representing, for each species, the sum of all branch lengths from the root to the tip. We simulated MVN using the *mvrnorm* function in the R package MASS (Venables

& Ripley, 2002). When using this auxiliary trait to impute target trait values, we expected that using the phylogeny into the imputation methods would not improve our analysis (i.e. provide no information on missing data) since the probability of missing values would only be correlated with the auxiliary trait and not with the phylogeny.

Missing data scenarios

To create missing data, we used the target trait simulated above and deleted different percentages of its values following three scenarios of missing data: Missing Completely at Random (MCAR), Missing at Random but phylogenetically structured (MAR.PHYLO), and Missing at Random but correlated with another phylogenetically unstructured trait (MAR.TRAIT). We created the MCAR scenario by randomly sampling a percentage (see below) of species along each phylogeny and replacing their trait values with missing values. For the MAR.PHYLO scenario, we sampled a species in each phylogeny and selected a percentage of its closest species to replace their trait values with missing values, allowing a strong missing data pattern that was phylogenetically structured. For the last scenario, MAR.TRAIT, we used the auxiliary trait (see above) to replace values in the target trait. We ordered the values of the auxiliary trait in ascending order and replaced the first percentage of values of the target trait with missing values. This represented a missing data pattern correlated with another trait, different to the target one. For each scenario, we simulated different percentages of missing values in the target trait: 5, 10, 20, 50, 70 and 90% of missing data. These percentages were chosen to represent common proportions of missing data present in highly used databases such as PanTHERIA (Jones et al., 2009) and EltonTraits (Wilman *et al.* 2014) (Fig. S1, Appendix S1).

Imputation methods

We evaluated four methods often applied by researchers to handle missing data: imputation based on averaging values (MEAN), no imputation and simply deleting missing values (LISTWISE), phylogenetic eigenvector maps (PEM), and multiple imputation by chained equations (MICE).

We used the MEAN method to impute missing values by filling them with the average of the observed values of the target trait. Under the LISTWISE method, we did not impute values but simply deleted those species with missing values in the phylogenies before the analyses. The PEM method uses both phylogenetic eigenvectors (Diniz-Filho *et al.* 1998) and traits to impute data considering different OU processes (Guénard et al. 2013). We applied this method in two ways: first, using only the phylogenetic eigenvectors (PEM.notrait) and, second, using these eigenvectors and the auxiliary trait (PEM.trait). By applying the PEM method in these two ways allowed us to evaluate whether phylogenetic information alone could impute data well or auxiliary traits were necessary. Eigenvector selection and fitting of trait evolutionary models were performed using the *MPSEM* R package (Guénard et al., 2013) using forward selection based on the second-order Akaike Information Criterion. The MICE method simulates several possible values for missing data from a posterior predictive distribution, then runs analysis and pools results over all simulated data (van Buuren et al., 2006). We chose this method because it is flexible and allows imputing categorical, continuous, and non-normally distributed data (van Buuren et al., 2006). We applied MICE by creating 10 datasets to run our analysis over them and pooled the results. The quantity of datasets created by MICE is dependent on the percentage of missing data and more datasets can provide higher accuracy and power in the analyses (Graham et al., 2007; Enders, 2010; van Buuren, 2012). However, because our objective was simply to estimate statistical bias

instead of inference power, 10 datasets can be considered appropriate (Graham et al., 2007). As with the PEM method, we applied MICE in two ways: only considering the auxiliary trait (MICE) and using this trait plus the phylogenetic eigenvectors selected as in PEM (MICE.phylo). We imputed data with MICE using the *mice* R package (Buuren & Groothuis-Oudshoorn, 2011).

We simulated 540 scenarios representing each combination of missing data percentage, mechanism, OU selection strength, and imputation methods. For each scenario, we simulated 100 replicates, thus producing 54000 independent results.

Estimating phylogenetic signal

We calculated the phylogenetic signal (PS) in our simulated phylogenies using two metrics: Blomberg's K (Blomberg et al., 2003) calculated with the *phylosig* function of *phytools* (Revell, 2012) and Moran's *I* correlograms (Moran's Correlogram) (Gittleman & Kot, 1990; Diniz-Filho, 2001). For calculating these correlograms, we created a phylogenetic distance matrix per phylogeny using the *cophenetic* function of *ape* (Paradis et al., 2004) and built the correlograms with the *lets.correl* function of the *letsR* R package (Vilela & Villalobos, 2015). Then, we used the Moran's I in the first distance class off the correlogram as indicative of PS, taking into account the non-linearity of correlograms generated under OU processes (Diniz-Filho, 2001).

Imputation effects on phylogenetic signal and descriptive statistics

Traditionally, performance evaluation of imputation methods have focused on common descriptive statistics such as (mean, variance, regression coefficient) (Collins et

al., 2001; van Buuren et al., 2006; Penone et al., 2014) instead of phylogenetic patterns. Therefore, we also evaluated the effect of imputed data on the estimation of such descriptive statistics. We calculated the mean and variance of the target trait as well as the regression coefficient (Ordinary Least Square) between the target trait and the auxiliary trait, before producing missing data and after imputing such data. Next, we measured the estimation error for these statistics as the squared error (SE), as below:

$$SE_i = (\tau 1_i - \tau 0_i)^2 \qquad \text{eqn2}$$

where $\tau 1$ represents the statistics calculated over imputed traits, $\tau 0$ is the statistics calculated from original traits.

Imputation error

To measure the potential error introduced by imputation methods, that is the deviation between imputed and original data, we followed Penone *et al.* (2014) and used the normalized root mean squared error (NRMSE):

$$NRMSE = \sqrt{\frac{mean((y - yimputed)^2)}{\max(y) - \min(y)}} \qquad \text{eqn 5}$$

where y is the original trait value, $y_{imputed}$ is the imputed value, max(y) and min(y) are the maximum and minimum values of the original trait, respectively. NRMSE varies between 0, no estimation error, and 1, maximum error (Oba et al., 2003).

Overall analyses

We were also interested on evaluating the effects of percentage of data missing, missing data mechanism, OU selection strength, and imputation methods as factors influencing the abovementioned effects of imputation. To do so, we grew regression trees (Hastie et al., 2009) with these factors as predictors and estimation errors, separately, as

individual responses. Nonetheless, as every simulation were stochastic, we accounted for this variance in our analysis by using as predictor variable the statistics calculated over original data ($\tau 0$). We grew regression trees using the *rpart* R package and variable importance was calculated as sum of improvements on the sum of squares in each node split by that variable. In addition, given concerns on the accuracy of imputation methods (Guénard et al., 2013; Penone et al., 2014), we also plotted the relationship between imputation error (NRSME) and estimation errors (SE) caused by imputation. All simulations and analysis were run in R 3.4.0 (R Core Team, 2017).

Results

In our simulations, we found that differences in estimation errors were dependent on missingness mechanism, imputation method, evolutionary model, percentage of missing data and original statistics, despite some differences in variable importance among statistics (Table.1).

Table.1. The importance of each variable to explain each estimated statistic errors. Variable importance were calculated as model improvement in each node split by the variable.

Statistic errors	Percentage	Alpha	Mechanisms	Methods	Original statistics
Blomberg's K	47.65	17.25	3.52	12.61	18.97
Moran's Correlogram	82.29	0	0	17.42	0.29
Mean	63.66	15.79	15.39	0	5.16
Variance	34.71	28.35	0	7.2	29.74
Regression coefficient	78.7	3.53	5.75	6.19	5.83

Moreover, each statistic had different hierarchical relationship among variables explaining their estimate errors, including the original value of the variable in some cases (Fig. S2-S6, Appendix S1). Besides, imputation errors showed different results between trait correlations (target vs. auxiliary trait; *r*) of 0.6 and 0.9, but descriptive statistics and phylogenetic signal errors did not show different results concerning this correlation. Therefore, we present all results for $r = 0.6$. Full results for $r = 0.9$ can be found in the Appendix S2.

Not surprisingly, our results showed a clear tendency of increasing error in estimating phylogenetic signal and descriptive statistics as the percentage of missing data gets larger (Fig. 2-4, Table.1). We did not identify a clear threshold in the amount of missing data that would guarantee lower statistical errors, but we identified two groups in the distribution of estimate errors split by a missing data percentage of 10 % (Fig.S2-S6, Appendix S1).

When data were missing completely at random (MCAR), most imputation methods showed good performance (Fig. 2-4; Fig. S7 and S8, in Appendix S1), except the MEAN method. Nevertheless, when estimating Blomberg's K, only LISTWISE showed low estimation errors (Fig. 2). For mean and regression coefficient estimation, imputation methods worked better when data were missing at random but correlated with another trait (MAR.TRAIT) than when data were missing and phylogenetically structured (MAR.PHYLO) (Fig. 4; Fig. S7 and S8, Appendix S1). Nevertheless, the Moran's Correlogram had low errors in all scenarios (Fig. 3).

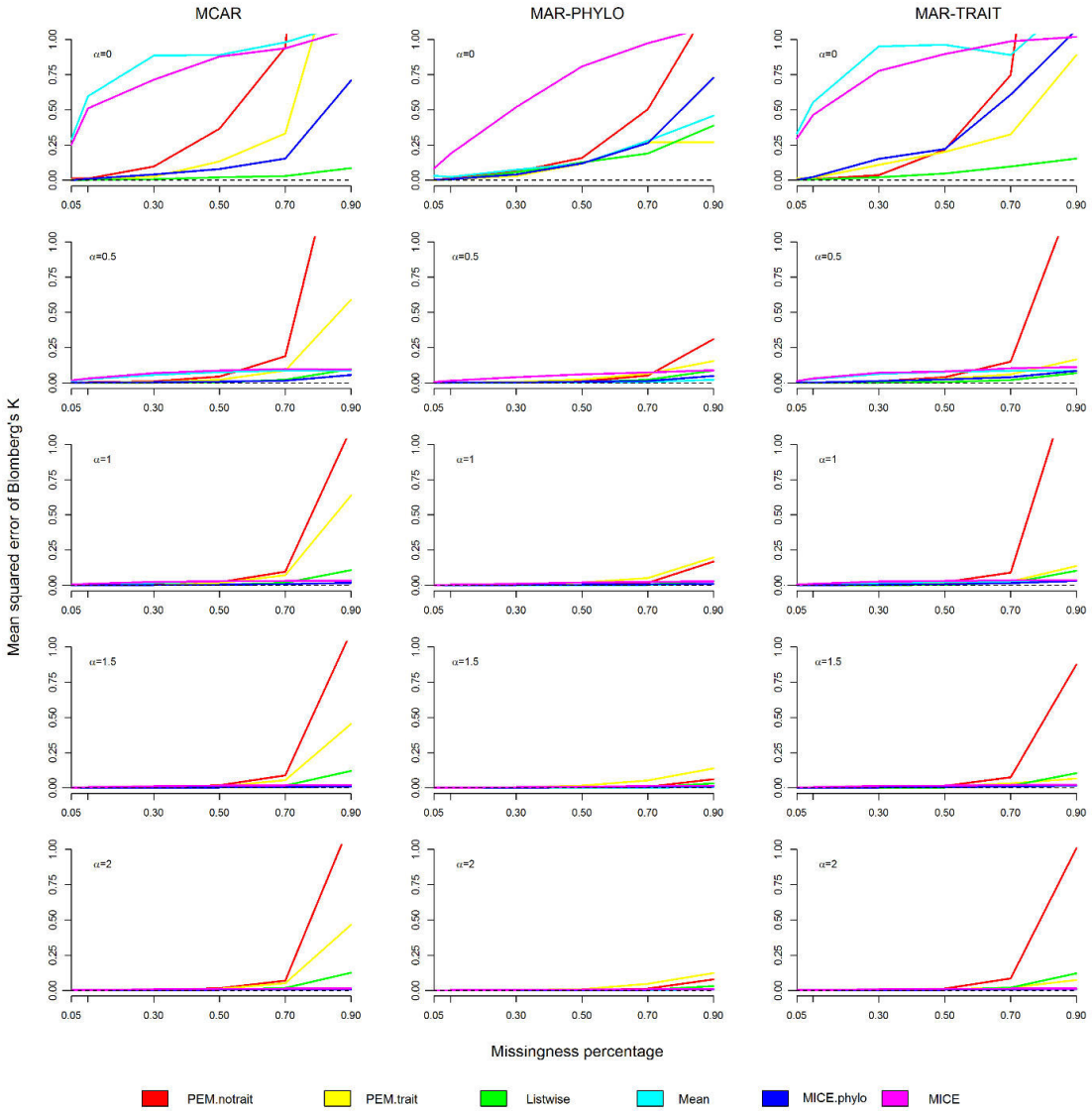


Figure 2. Mean squared error of Blomberg's K under different methods, OU selective strength, missing data percentage and mechanisms.

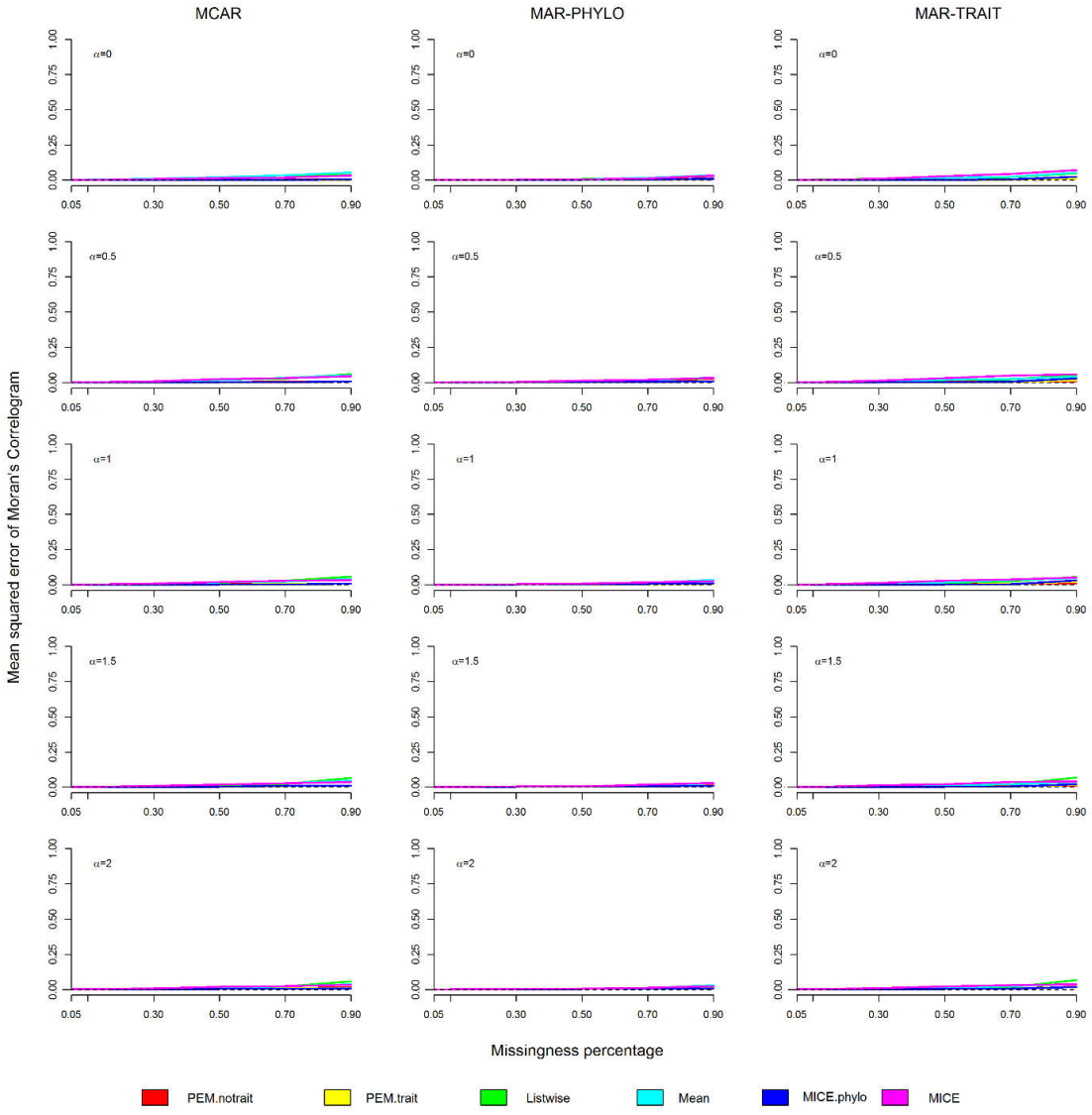


Figure 3. Mean squared error of Moran's Correlogram under different methods, OU selective strength, missing data percentage and mechanisms.

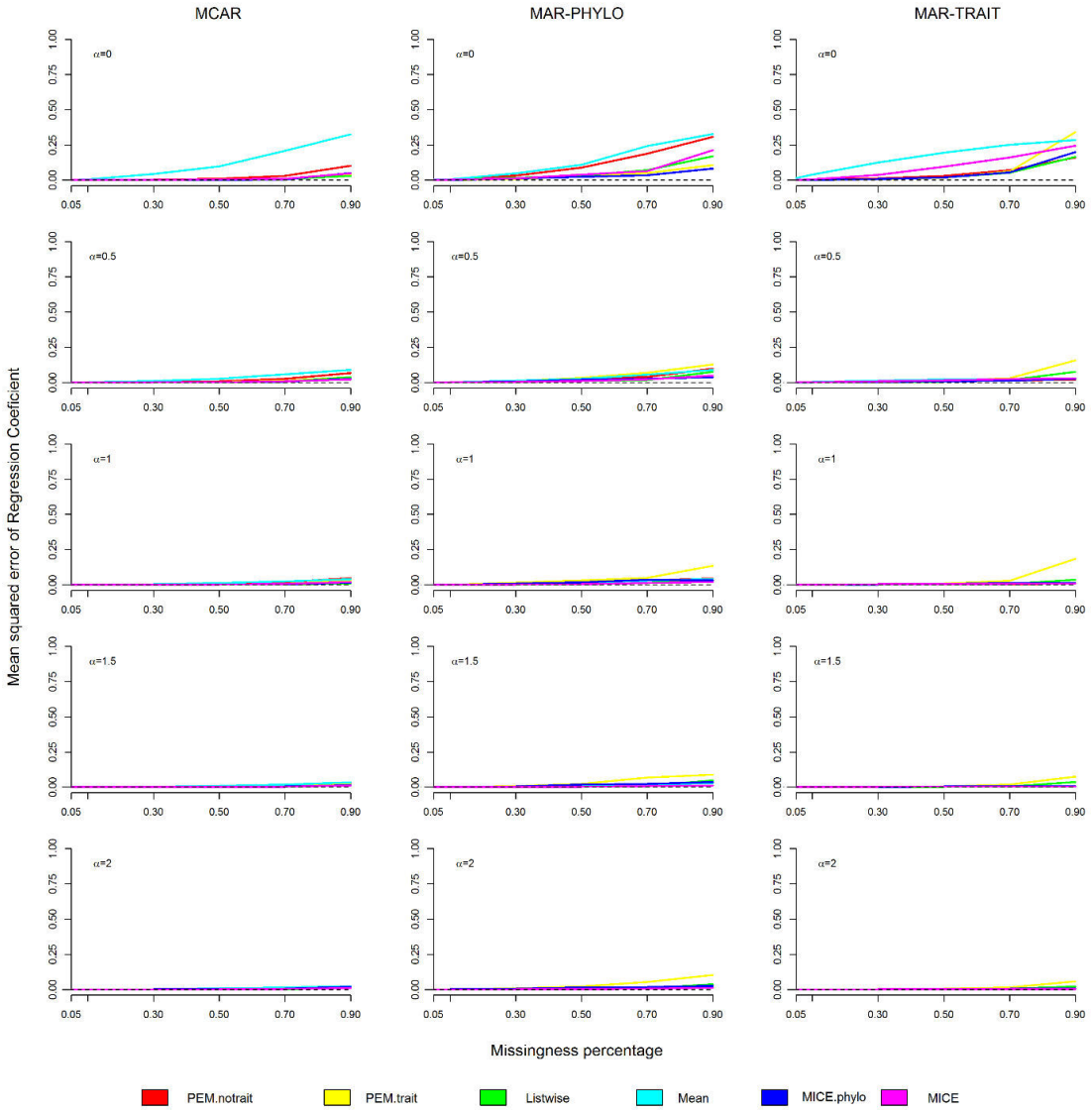


Figure 4. Mean squared error of regression coefficient under different methods, OU selective strength, missing data percentage and mechanisms.

The level of selection strength on trait evolution under the OU process was important just to explain SE and Blomberg's K (Table.1). Nonetheless, we found a tendency to SE to decrease as the selection strength increased from pure evolutionary drift

(i.e. OU alpha = 0; Brownian motion) to strong selection (OU alpha = 2) (Fig. 2-4; Fig. S7 and S8, Appendix S1).

The less sensitive methods were those that considered phylogenetic information in the imputation process (Fig. 2-5). PEM.trait, PEM.notrait, and MICE.phylo showed results less sensitive over different mechanisms of missing data (Fig. 2-4; Fig. S7 and S8, Appendix S1). The MEAN method was the most sensitive (Fig. 2-4; Fig. S7 and S8, Appendix S1).

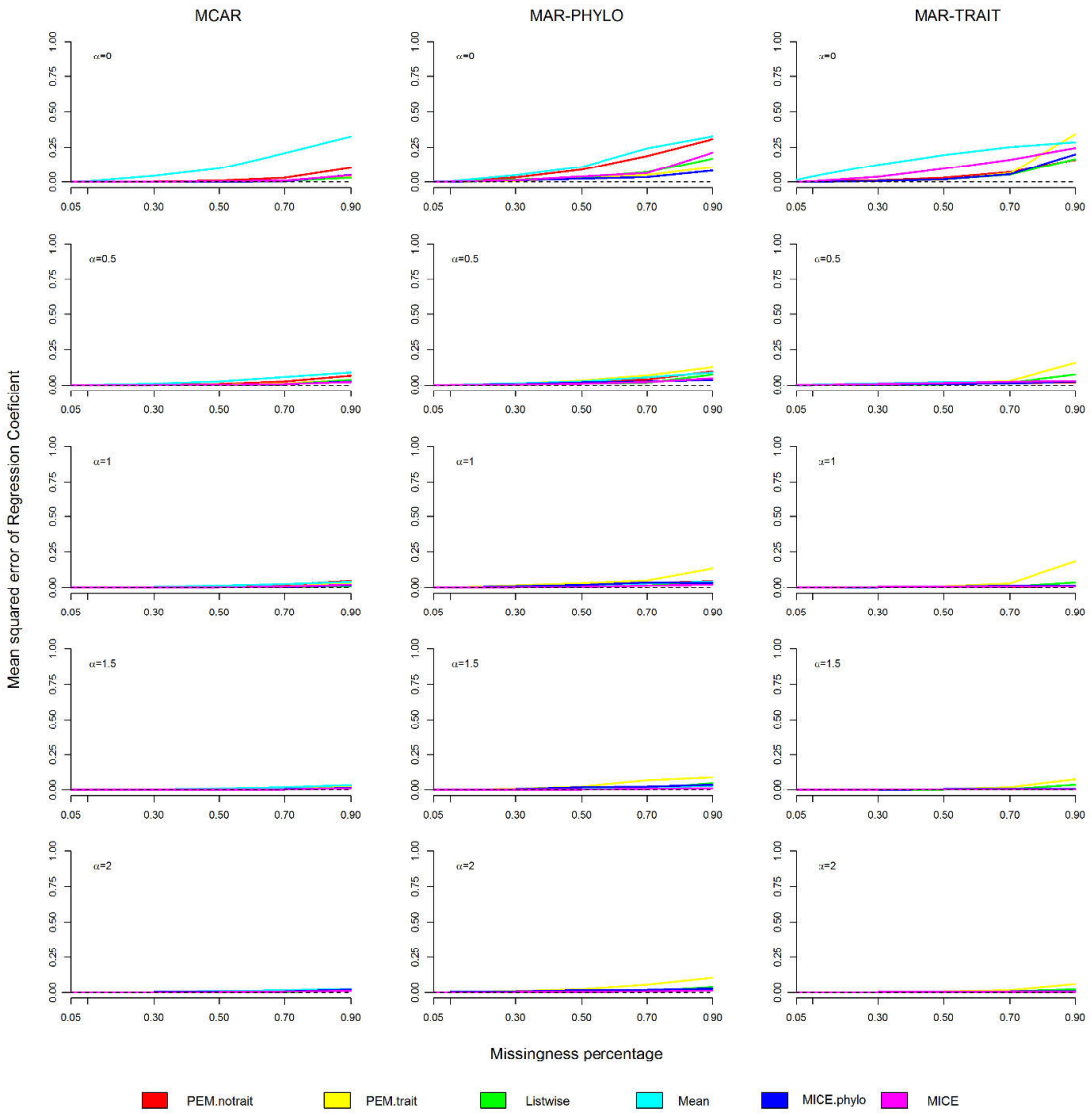


Figure 5. Change percentage of Blomberg's K under different methods, OU selective strength, missing data percentage and mechanisms. Change percentage was calculated as imputed Blomberg's K minus original Blomberg's K divided by original Blomberg's K.

Phylogenetic signal metrics (Blomberg's K and Moran's Correlogram) were lower than the original (before imputation) when using MEAN and MICE methods (Fig. 5 and 6). All other methods estimated Moran's Correlogram correctly under most simulated scenarios (Fig. 6), whereas the estimation of Blomberg's K showed different patterns. Blomberg's K was overestimated by PEM.trait and PEM.notrait and underestimated by MICE, even under the MCAR missing mechanism (Fig. 5). Nevertheless, Blomberg's K estimation errors decreased when phylogenetic eigenvectors were used in MICE.phylo (Fig. 5).

Descriptive statistics (mean and regression coefficient) were well estimated by all imputation methods (except MEAN) under MCAR. MAR.TRAIT and MAR-PHYLO generated biased estimations, but these biases were higher under MAR-PHYLO (Fig. 4, Fig. S7 and S8, Appendix S1). Nonetheless, variance had high estimations errors in all mechanisms, independent of the imputation methods (Fig. S8, Appendix S1)

Finally, we found that imputation errors were correlated with estimate errors (SE) (Fig. 7). In addition, the imputation error and estimate error relationship evaluated here was asymptotic in log-scale, thus as imputation error increases the estimation error increases faster.

Discussion

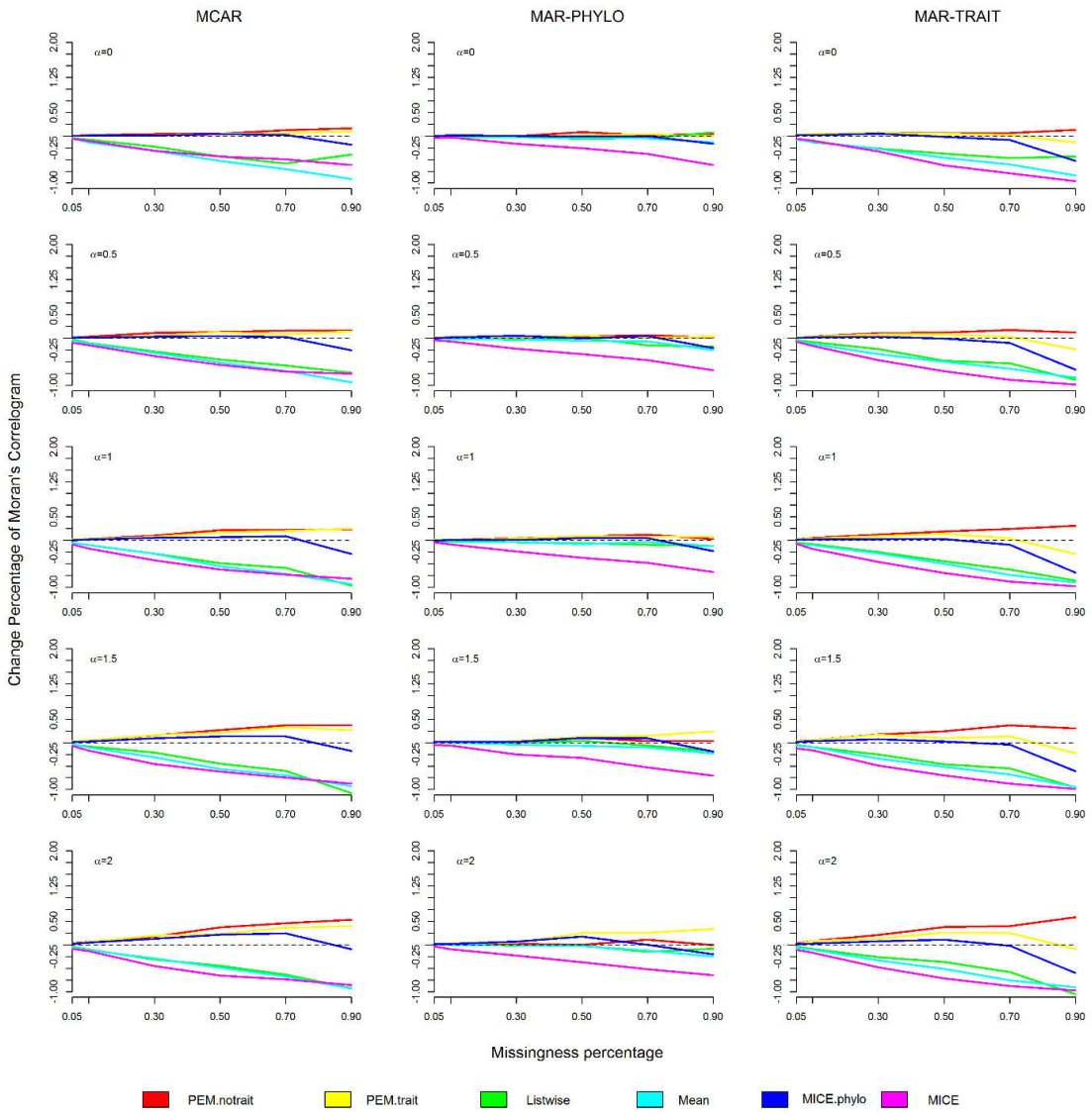


Figure 6. Change percentage of Moran's Correlogram under different methods, OU selective strength, missing data percentage and mechanisms. Change percentage was calculated as imputed Moran's Correlogram minus original Moran's Correlogram divided by original Moran's Correlogram.

Ecologists and evolutionary biologists are increasingly creating, using, and sharing large trait databases that are inevitably sparse and often completed by imputing missing values (Guénard et al., 2013; Swenson, 2014; Schrodt et al., 2015). Here we argue

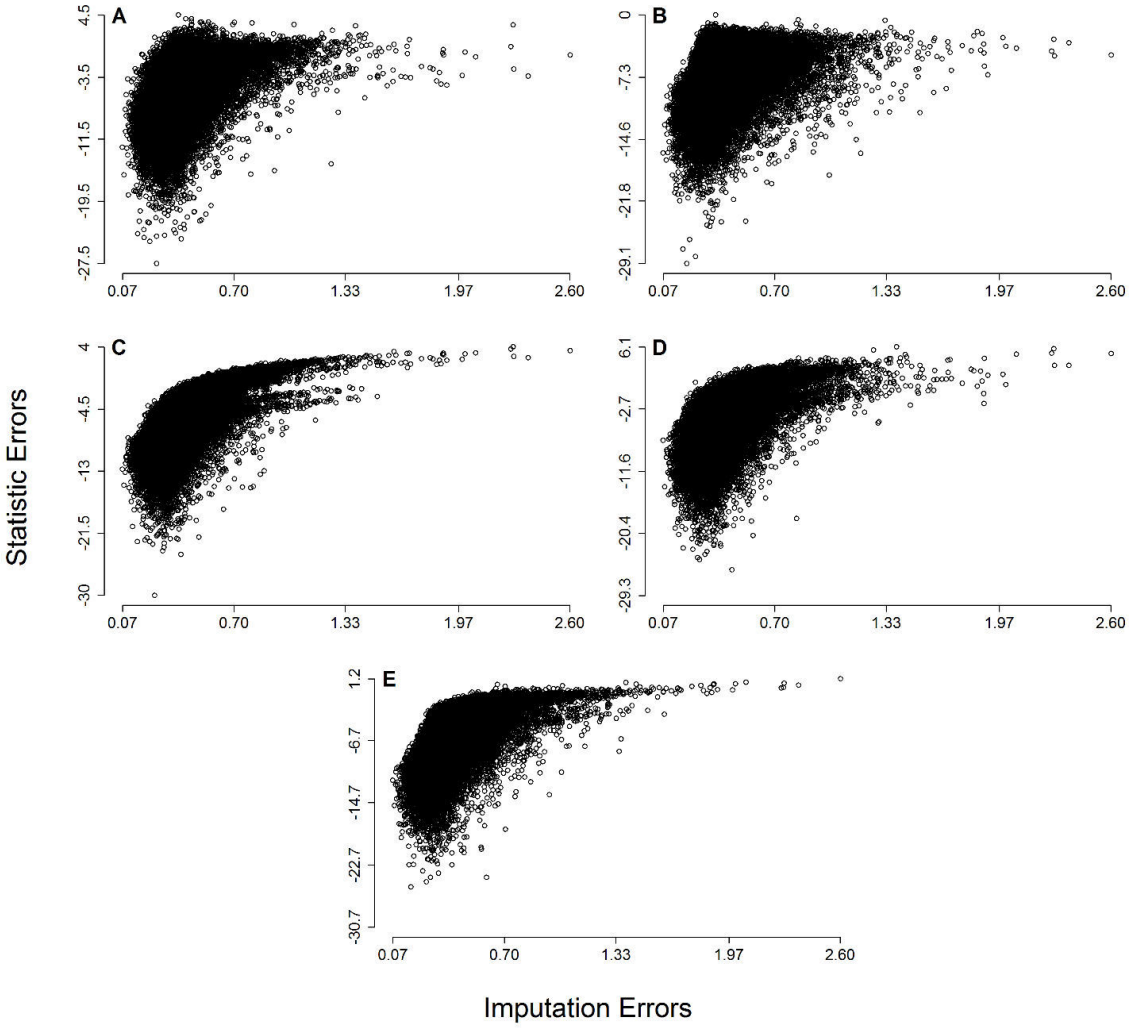


Figure 7. Scatterplot of imputation errors (average NRMSE) and statistical errors. (A) Logarithm of Blomberg's K SE (squared error), (B) Logarithm of Moran's Correlogram SE, (C) Logarithm of Mean SE , (D) Logarithm of Variance SE and (E) Logarithm of Regression coefficient SE.

that we should be extremely careful when using imputed databases, even for the estimation of simple parameters (i.e. means, variances and regression coefficients). Our findings revealed that estimations based on imputed data depends on every aspect of data

property and strategy of analysis, as percentage of missing data, source/mechanism of absence, trait evolution, methods for gap filling, and statistics or parameters to be estimated (Fig.8). This has commonly been acknowledged in statistical research (Rubin, 1976; Enders, 2010) and should begin to be so in the ecological and evolutionary research as claimed by Nakagawa & Freckleton (2008). Based on our results, we can infer that the large changes in the estimations, due to different analytical choices, may also be an important cause of irreproducibility in our field (Borregaard & Hart, 2016).

The most pervasive obstacle for deriving conclusions from large datasets is simply the proportion of those species lacking data. Previous studies found that reliable estimations from imputed data can be made when up to 60% of the values were missing (Barzi, 2004; Penone et al., 2014). However, in our results, the effect of missing data percentage was not direct, but rather interacted with all of the other aspects evaluated here. Thus, there is no simple way of deriving a threshold on how much missing data would be allowed to be imputed and still make reliable estimations.

Knowing the causes of data absence is the first issue to be sorted out before any analysis (van Buuren 2012). The most common assumption in ecological and evolutionary studies is that data is missing completely at random (MCAR). This is evident in the wide variety of functions of the most commonly used software (the R programming language) allowing deleting missing values indiscriminately. Indeed, if data were under MCAR, previous findings and ours showed that estimations based on deletions and imputations could safely be made (Nakagawa & Freckleton, 2010; Penone et al., 2014; Taugourdeau et al., 2014). However, biological data are rarely missing completely at random (Nakagawa & Freckleton, 2008; Enders, 2010). For instance, bias in ecological data absence can be related to the fact that some taxa are most studied than others (Gonzalez-Suarez *et al.* 2012). Moreover, such bias can stem from body mass differences

among species, where large species have a higher probability of being described first (Vilela et al., 2014) and have their data collected (Gonzalez-Suarez et al., 2012) compared to small species. Also, species present in easily accessible regions are better studied than those occurring in regions that are hard to access (Reddy & Dávalos, 2003). In our simulations, higher biased estimates were found when data were missing at random but correlated with other variable (MAR), especially phylogeny (MAR.PHYLO). Such results differ from those found by Penone et al. (2014), who did not find significant estimation differences among missing data mechanisms. This discrepancy could be related to our way of simulating MAR.PHYLO, creating a stronger phylogenetic structure than that simulated by them.

Our simulations revealed that imputation methods considering phylogenetic structure (PEM.trait, PEM.notrait and MICE.phylo) performed better than methods not doing so (MEAN, LISTWISE, and MICE) under all missing data mechanisms (MCAR, MAR.PHYLO, and MAR.TRAIT). Such findings support previous claims favoring "phylogenetic imputation" as a powerful tool in predicting missing species values (Penone *et al.*, 2014; Swenson 2014). More interestingly, our results showed that some phylogenetic imputation methods (PEM.notrait) perform better than non-phylogenetic ones, even when missing data was uncorrelated with phylogeny but to an auxiliary trait (MAR.TRAIT). This result was unexpected based on missing data theory, which suggests that under MAR.TRAIT some variable correlated with missing data probability is required to guarantee reliable estimations (Enders 2010).

Overall, MICE.phy, PEM.notrait and PEM.trait performed best among all imputation methods tested. However, they committed high estimation error in MAR mechanisms, what we did not expected previously. We do not know if this performance is an effect of eigenvectors selections, what was performed over missing values and after

used to impute missing values. The selection of eigenvectors over traits with missing data could result in misrepresentation of phylogenetic structure.

PEM.notrait and PEM.trait represented single imputation method, which imputes a single value for each missing datum, thus not accounting for uncertainty of the imputed value. Consequently, PEM methods may underestimate standard errors and bias subsequent hypothesis testing (i.e. increasing Type I error rates) (Enders 2010; van Buuren 2012). To avoid such biases, the statistical literature suggests using multiple imputation methods (Schafer & Graham, 2002; Enders, 2010; van Buuren, 2012), as we represented by MICE methods. However, our results did not show better performance of MICE, even when including phylogenetic information, in estimating descriptive statistics or phylogenetic signal compared to PEM. Despite multiple imputation being one of the most suggested methods for handling missing data (van Buuren 2012), additional research is necessary to evaluate its performance with phylogenetically structured data.

Filling missing values by averaging the observed ones (MEAN) or simply deleting species with missing values (LISTWISE) generated poor estimates, which is related to the fact that both methods assume that data is MCAR. MEAN only worked satisfactorily for estimating the trait average. LISTWISE disrupts the distribution of trait values, thus results in biased estimates (Enders, 2010). However, this method performed well when estimating phylogenetic signal. This is encouraging, given that researchers interested in trait phylogenetic signal usually delete missing values (Blomberg & Garland, 2002; Kamilar & Cooper, 2013) thus guaranteeing potentially unbiased results.

Phylogenetic imputation is based on the assumption of target traits being phylogenetically structured (i.e. showing phylogenetic signal; Swenson 2014). However, phylogenetic structure is dependent on how traits evolved (Diniz-Filho, 2001; Guénard et al., 2013). Accordingly, trait evolution was an important issue in our study. Across our

simulated scenarios, estimation errors were higher when target traits were simulated under Brownian motion (BM) than under OU processes, agreeing with previous study (Guénard et al., 2013). Better estimates under OU than BM processes may result from higher trait resemblance and lower variance among species generated when increasing selection strength under OU processes (Hansen, 1997; Butler & King, 2004). Thus, predicting missing values of target traits will benefit from knowing their particular evolutionary model and will be more accurate if such traits evolved under strong selection regimes. Again, this suggests that researchers need to find the appropriate evolutionary model for their target traits before judging the need to use phylogenetic imputation methods for handling missing data. It should be noted, however, that fitting evolutionary models over incomplete data could itself be biased owing to the use of observed values only and thus pruned phylogenies (Slater et al., 2012).

Phylogenetic imputation methods may also produce bias when estimating phylogenetic signal. More specifically, our findings suggest that such methods can actually alter the original phylogenetic structure of the trait (i.e. the structure if data were complete). In fact, PS may be incorrectly estimated even under MCAR. Moreover, when using Blomberg's K, imputation by PEM overestimated the original phylogenetic signal of the target trait (i.e. created when the trait was simulated) whereas MICE.phylo underestimated it.

In addition, PS estimation errors were dependent on the evaluated metric. Regardless of the simulated scenario, estimation errors were lower for PS based on Moran's I correlogram than Blomberg's K. Similarly, Münkemüller et al. (2012) showed that Moran's I is less sensible than Blomberg's K to changes in trait phylogenetic structure even when random noise is added. Blomberg's K measures a global pattern along a phylogeny, based on observed and expected total trait variance under Brownian

motion (Blomberg et al., 2003), whereas Moran's I correlogram measures the correlation of trait values within different phylogenetic distance classes (Gittleman & Kot, 1990). Therefore, changes in total trait variance caused by imputation may not have strong impacts on within-first class correlations, rendering Blomberg's K more sensitive than Moran's Correlogram to such changes.

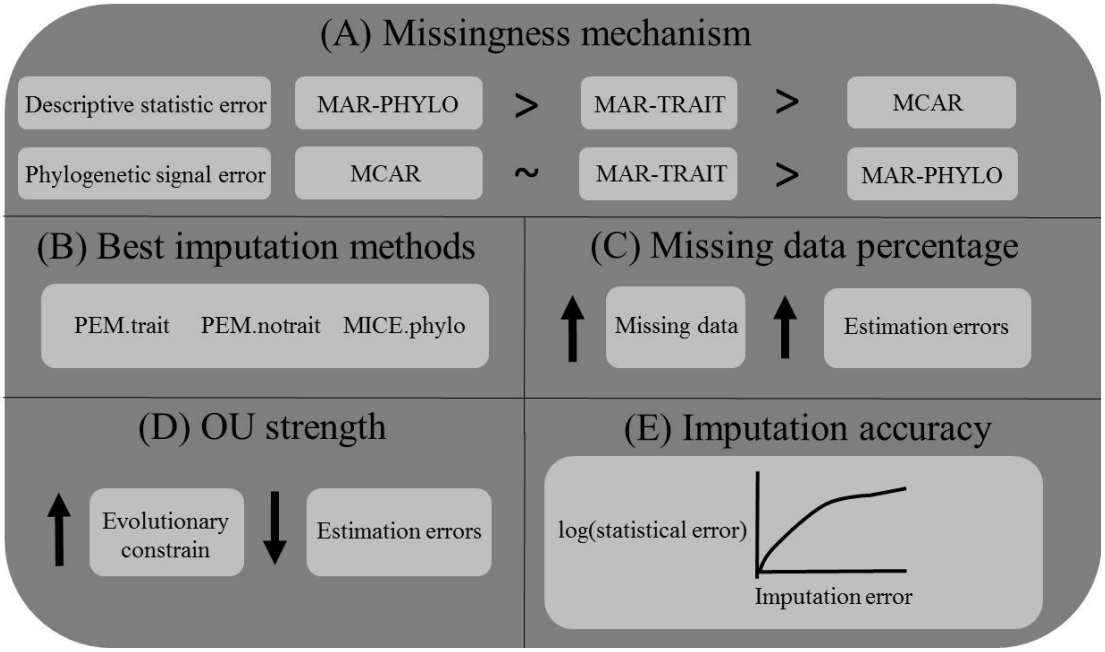


Figure 8 Summary of the main results showing (A) the differences on estimation errors among missing data mechanisms and estimated statistics; (B) highlighting the best imputation methods; (C) the effect of missing data percentage in statistical estimation; (D) OU selection strength; and (E) the non-linear relationship between imputation error and statistical estimation error logarithm.

New proposed methods to fill sparse databases currently concerns about their degree of imputation error, that is how much imputed values deviate from the original trait values (Guénard et al., 2013; Penone et al., 2014; Schrodt et al., 2015). We found

that single and multiple phylogenetic imputation methods can be highly accurate, resulting in small deviations between imputed and observed values, as suggested by other authors (Guénard et al., 2013; Penone et al., 2014; Diniz-Filho et al., 2015; Schrodt et al., 2015). In addition, we found that imputation error was positively correlated with estimation errors but their relationship was not linear. That is, increasing imputation error causes estimation errors to increase much more rapidly. This is particularly relevant if researchers were to use imputed databases blindly –without correctly treating imputed values. Such practice could create spurious results. This is because even if imputation is accurate, imputed values simply represent one among several possibilities without providing information on imputation uncertainty. In fact, using an accurately imputed database does not necessarily mean that the original trait distribution and its relationship with other variables will be recovered (van Buuren 2012).

Concluding remarks

Instead of providing imputed trait databases, we should focus on treating missing values with appropriate methods. We have shown here that such methods should consider phylogenetic information. With the increase of computational literacy among ecologists and evolutionary biologists (Ram, 2013), we encourage researchers to use simulations of their data and methods to find the appropriate solution for their study goals. Furthermore, researchers need to develop phylogenetic methods that consider imputation uncertainty and preserve the original data's phylogenetic signal. Missing data is one of the most pervasive features of trait databases and the only effective solution for this Raunkiaeran shortfall is collecting more data. Nevertheless, acknowledging such shortfall instead of ignoring it will effectively help guiding research towards solving it.

Acknowledgements

We thank G. Guénard for advices on PEM and L. Revell for help on evolutionary simulations. L.J. was supported by a CAPES doctoral fellowship. FV was supported by a BJT "Science without borders" CNPq grant. L.M.B. and J.A.F.D.F. are continuously supported by CNPq productivity grants. We do not have a conflict of interest to declare.

References

Barzi, F. 2004. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *Am. J. Epidemiol.* **160**: 34–45.

Blomberg, S.P. & Garland, T. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *J. Evol. Biol.* **15**: 899–910.

Blomberg, S.P., Garland, T. & Ives, A.R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution.* **57**: 717–745.

Borregaard, M.K. & Hart, E.M. 2016. Towards a more reproducible ecology. *Ecography.* **39**: 349–353.

Butler, M.A. & King, A.A. 2004. Phylogenetic comparative analysis : a modeling approach for adaptive evolution. *Am. Nat.* **164**: 683–695.

Buuren, S. van & Groothuis-Oudshoorn, K. 2011. mice : Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45**.

Cavender-Bares, J., Kozak, K.H., Fine, P.V.A. & Kembel, S.W. 2009. The merging of community ecology and phylogenetic biology. *Ecol. Lett.* **12**: 693–715.

Collins, L.M., Schafer, J.L. & Kam, C.M. 2001. A comparision of inclusive and

restrictive strategies in modern missing data procedures. *Psychol. Methods* **6**: 330–351.

Cooper, N., Thomas, G.H., Venditti, C., Meade, A. & Freckleton, R.P. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biol. J. Linn. Soc.* **118**: 64–77.

Davis, M.P., Midford, P.E. & Maddison, W. 2013. Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. *BMC Evol. Biol.* **13**: 38.

Diniz-Filho, J.A.F. 2001. Phylogenetic autocorrelation under distinct evolutionary process. *Evolution.* **55**: 1104–1109.

Diniz-Filho, J.A.F., Sant'Ana, C.E.R. & Bini, L.M. 1998. An Eigenvector Method for estimating Phylogenetic Inertia. *Evolution.* **52**: 1247–1262.

Diniz-Filho, J.A.F., Villalobos, F. & Bini, L.M. 2015. The best of both worlds : Phylogenetic eigenvector regression and mapping. *Genet. Mol. Biol.* **38**: 396–400.

Enders, C.K. 2010. *Applied Missing Data Analysis*, 1st ed. New York, NY.

Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.* **125**: 1–15.

Freckleton, R.P., Harvey, P.H. & Pagel, M. 2002. Phylogenetic analysis and comparative data : a test and review of evidence. *Am. Nat.* **160**: 712–726.

Freckleton, R.P. & Jetz, W. 2009. Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. *Proc. R. Soc. B* **276**: 21–30.

Gaston, K.J., Chown, S.L. & Evans, K.L. 2008. Ecogeographical rules: elements of a synthesis. *J. Biogeogr.* **35**: 483–500.

Gillespie, D. 1996. Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. *Phys. Rev. E* **54**: 2084–2091.

Gittleman, J.L. & Kot, M. 1990. Adaptation: statistics and a null model for estimating phylogenetic effects. *Syst. Zool.* **39**: 227–241.

Goldberg, E.E., Kohn, J.R., Lande, R., Robertson, K. a., Smith, S. a. & Igic, B. 2010. Species selection maintains self-incompatibility. *Science.* **330**: 493–495.

Gonzalez-Suarez, M., Lucas, P.M. & Revilla, E. 2012. Biases in comparative analyses of extinction risk: mind the gap. *J. Anim. Ecol.* **81**: 1211–22.

Graham, J.W., Olchowski, A.E. & Gilreath, T.D. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* **8**: 206–213.

Guénard, G., Legendre, P. & Peres-Neto, P. 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. *Methods Ecol. Evol.* **4**: 1120–1131.

Hansen, T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution.* **51**: 1341–1351.

Hansen, T.F. & Martins, E.P. 1996. Translating between microevolutionary process and macroevolutionary patterns: correlation structure of interspecific data. *Evolution.* **50**: 1404–1417.

Harmon, L.J., Losos, J.B., Jonathan Davies, T., Gillespie, R.G., Gittleman, J.L., Bryan Jennings, W., *et al.* 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution.* **64**: 2385–2396.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning*. Springer New York, New York, NY.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **46**: 523–549.

Jetz, W. & Freckleton, R.P. 2015. Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**: 20140016.

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S. a., O'Dell, J., Orme, C.D.L., *et al.* 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**: 2648–2648.

Kamilar, J.M. & Cooper, N. 2013. Phylogenetic signal in primate behaviour, ecology and life history. *Proceeding R. Soc. B* **368**: 20120341.

Kattge, J., Ogle, K., Bönisch, G., Díaz, S., Lavorel, S., Madin, J., *et al.* 2011. A generic structure for plant trait databases. *Methods Ecol. Evol.* **2**: 202–213.

Lukas, D. & Clutton-Brock, T.H. 2013. The evolution of social monogamy in mammals. *Science.* **341**: 526–30.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., *et al.* 2012. How to measure and test phylogenetic signal. *Methods Ecol. Evol.* **3**: 743–756.

Nakagawa, S. & Freckleton, R.P. 2008. Missing inaction: the dangers of ignoring missing data. *Trends Ecol. Evol.* **23**: 592–596.

Nakagawa, S. & Freckleton, R.P. 2010. Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behav. Ecol. Sociobiol.* **65**: 103–116.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. & Ishii, S. 2003. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **19**: 2088–2096.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

Penone, C., Davidson, A.D., Shoemaker, K.T., Marco, M. Di, Rondinini, C., Brooks, T.M., *et al.* 2014. Imputation of missing data in life-history traits datasets: which approach performs the best? *Methods Ecol. Evol.* **5**: 961–970.

Purvis, A., Gittleman, J.L., Cowlishaw, G. & Mace, G.M. 2000. Predicting extinction risk in declining species. *Proceeding R. Soc. B* **267**: 1947–1952.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput.* R Foundation for Statistical Computing, Vienna, Austria.

Ram, K. 2013. Git can facilitate greater reproducibility and increased transparency in science. *Source Code Biol. Med.* **8**: 7.

Reddy, S. & Dávalos, L.M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *J. Biogeogr.* **30**: 1719–1727.

Reichman, O.J., Jones, M.B. & Schildhauer, M.P. 2011. Challenges and Opportunities of Open Data in Ecology. *Science.* **331**: 703–705.

Revell, L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**: 217–223.

Revell, L.J., Harmon, L.J. & Collar, D.C. 2008. Phylogenetic signal, evolutionary process, and rate. *Syst. Biol.* **57**: 591–601.

Rosado, B.H.P., de S. L. Figueiredo, M., de Mattos, E.A. & Grelle, C.E. V. 2015. Eltonian shortfall due to the Grinnellian view: functional ecology between the mismatch of niche concepts. *Ecography.* **39**: 1034–1041.

Rubin, D.. 1976. Inference and Missing Data. *Biometrika* **63**: 581–592.

Schafer, J.L. & Graham, J.W. 2002. Missing Data: our view of the state of the art. *Psychol. Methods* **7**: 147–177.

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., *et al.* 2015. BHPMF - a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Glob. Ecol. Biogeogr.* **24**: 1510–1521.

Slater, G.J., Harmon, L.J., Wegmann, D., Joyce, P., Revell, L.J. & Alfaro, M.E. 2012. Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate bayesian computation. *Evolution.* **66**: 752–762.

Swenson, N.G. 2014. Phylogenetic imputation of plant functional trait databases. *Ecography.* **37**: 105–110.

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O. & Amiaud, B. 2014. Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data. *Ecol. Evol.* **4**: 944–958.

van Buuren, S. 2012. *Flexible Imputation of Missing Data*, 1st ed. Chapman and Hall/CRC, Boca Raton, Fl.

van Buuren, S., Brands, J.P.L., Groothuis-Oudshoorn, K. & Rubin, D.B. 2006. Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* **76**: 1049–1064.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*, 4th ed.

Springer, New York.

Vilela, B. & Villalobos, F. 2015. letsR: a new R package for data handling and analysis in macroecology. *Methods Ecol. Evol.* **6**: 1229–1234.

Vilela, B., Villalobos, F., Rodríguez, M.Á. & Terribile, L.C. 2014. Body Size, extinction risk and knowledge bias in New World snakes. *PLoS One* **9**: e113429.

Webb, C.O., Ackerly, D.D., Mcpeek, M.A. & Donoghue, M.J. 2002. Phylogenies and community ecology. *Annu. Rev. Ecol. Evol. Syst.* **33**: 475–505.

Wiens, J.J. & Graham, C.H. 2005. Niche Conservatism: integrating evolution, ecology, and conservation biology. *Annu. Rev. Ecol. Evol. Syst.* **36**: 519–539.

Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M.M. & Jetz, W. 2014. EltonTraits 1.0 : Species-level foraging attributes of the world ’s birds and mammals. *Ecology* **95**: 2027.

APPENDIX 1

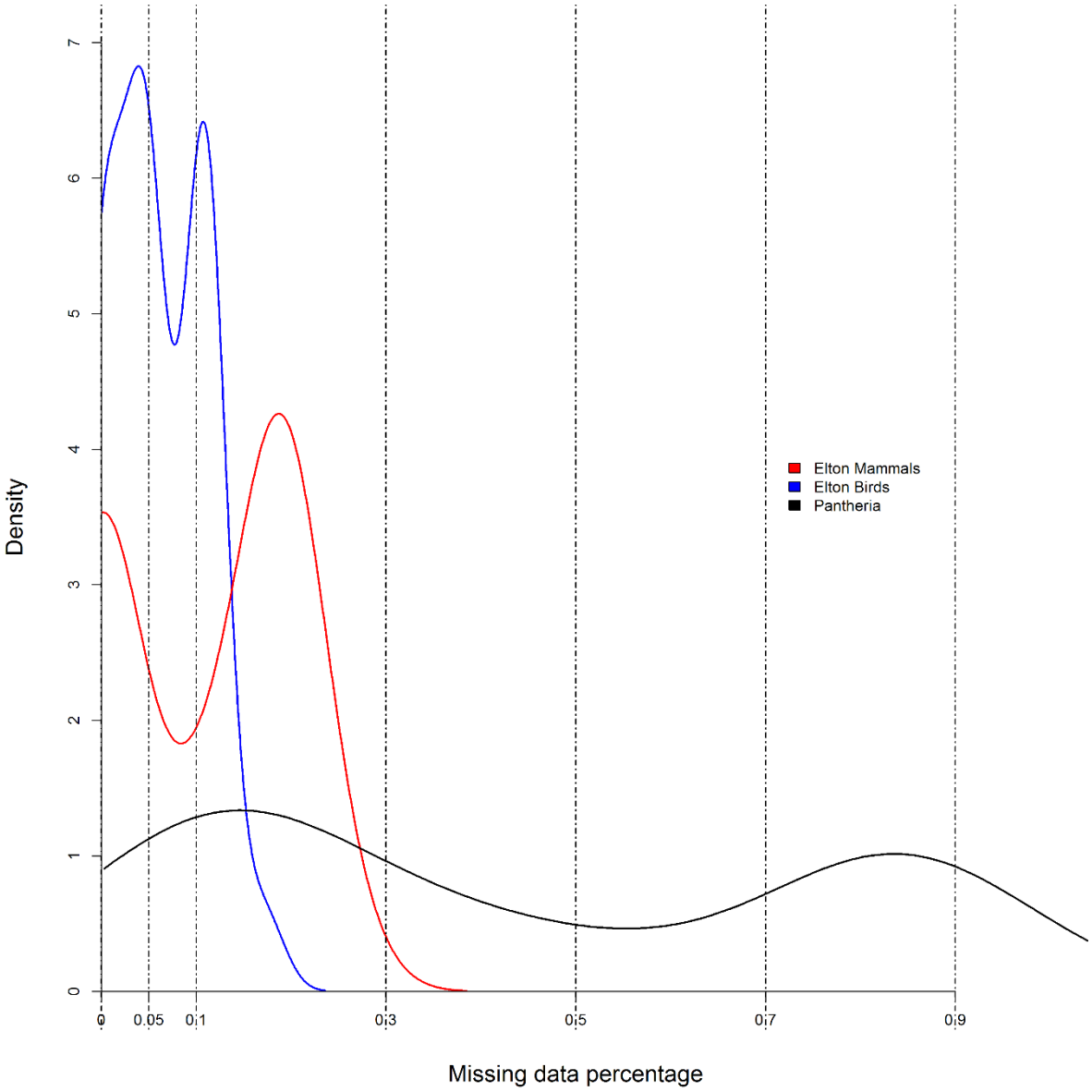


Figure S1. Density of missing values percentage in each database. Dotted lines represents missingness percentage simulated. Colored lines represent density of percentages through different traits in each database. Low percentage represents EltonTraits missing data percentage, high values represents Pantheria missing data percentage and intermediate represents possible values to be found in other database.

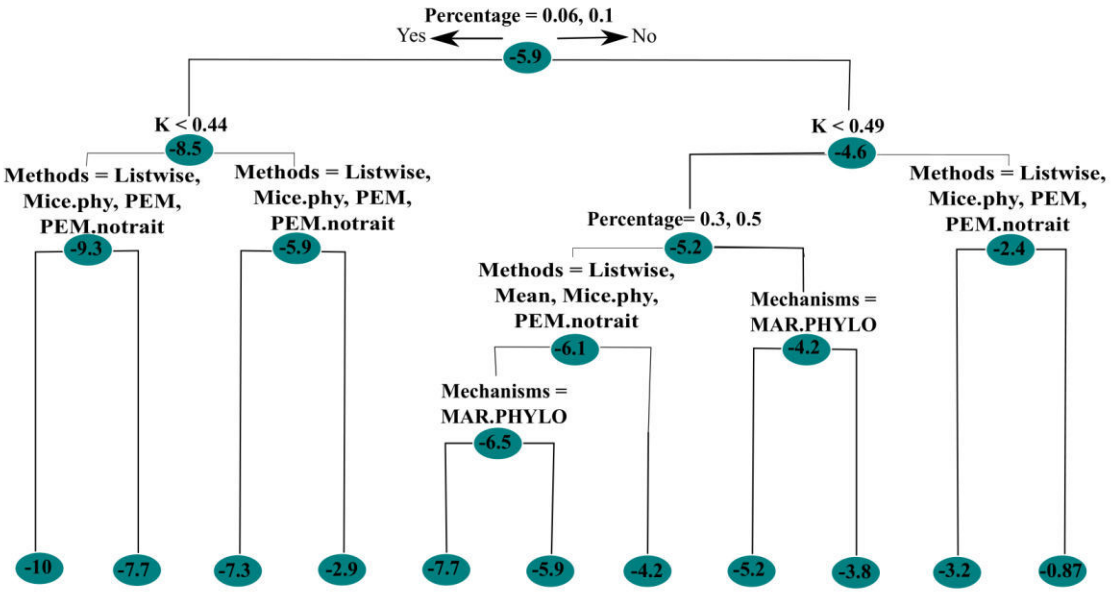


Fig S2. Regression tree analysis of the logarithm of Blomberg's K squared error (SE). Each node contains the mean of the logarithm SE in each tree split. In the nodes, there are the criterion and the variables used in each split.

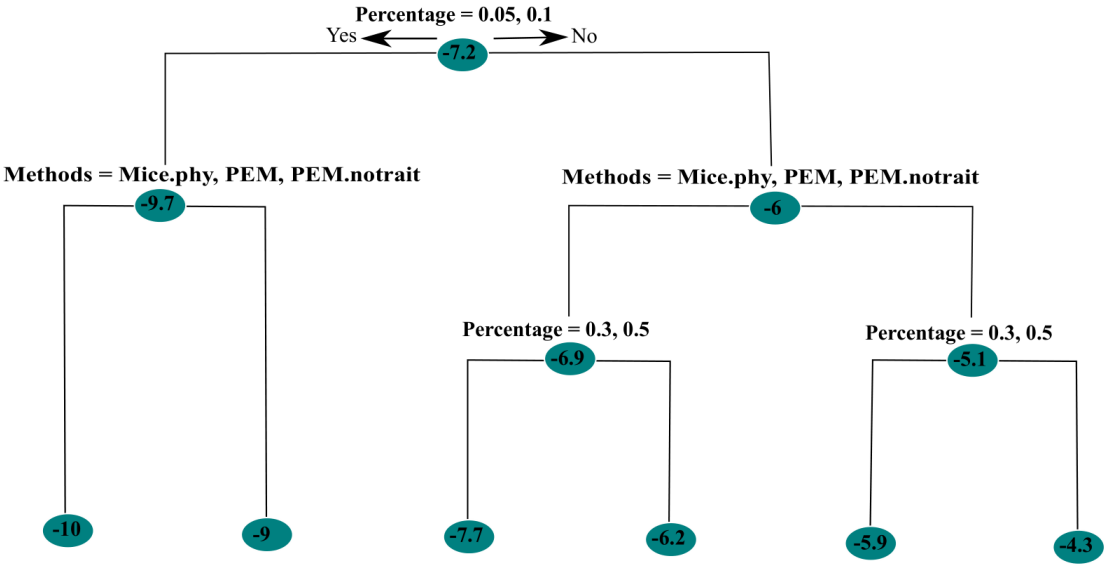


Fig S3. Regression tree analysis of the logarithm of Moran's Correlogram squared error (SE). Each node contains the mean of the logarithm SE in each tree split. In the nodes, there are the criterion and the variables used in each split.

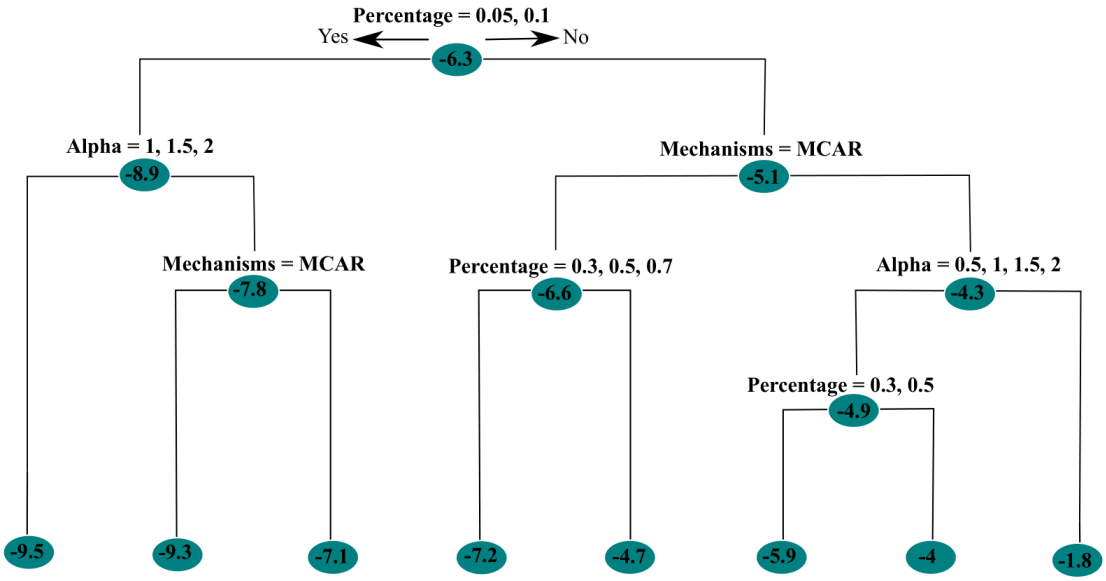


Fig S4. Regression tree analysis of the logarithm of squared error of mean (SE). Each node contains the mean of the logarithm SE in each tree split. In the nodes, there are the criterion and the variables used in each split.

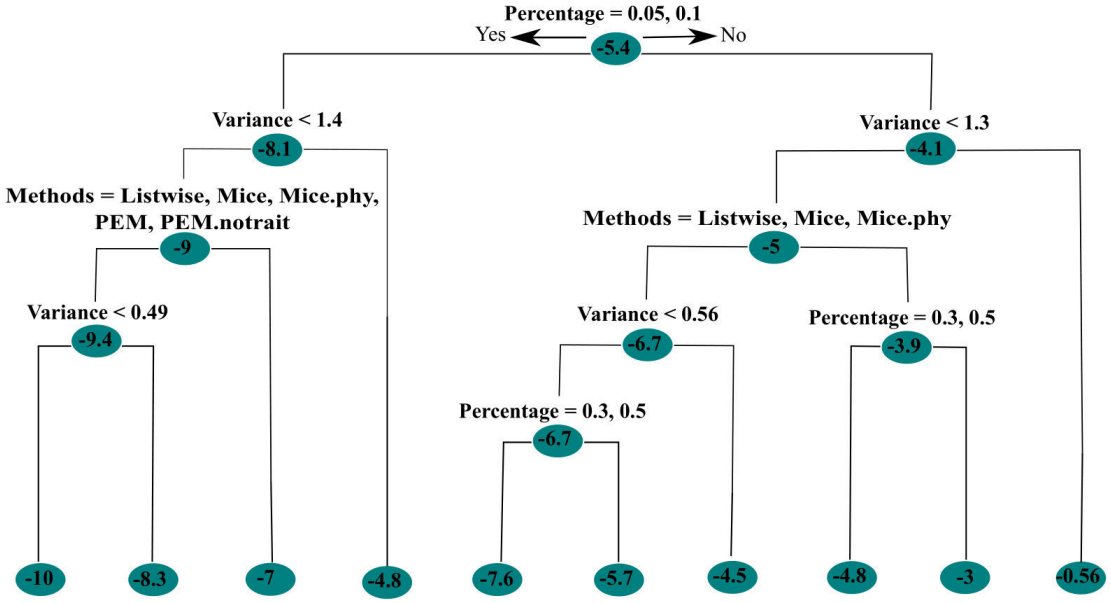


Fig S5. Regression tree analysis of the logarithm of trait's variance squared error (SE). Each node contains the mean of the logarithm SE in each tree split. In the nodes, there are the criterion and the variables used in each split.

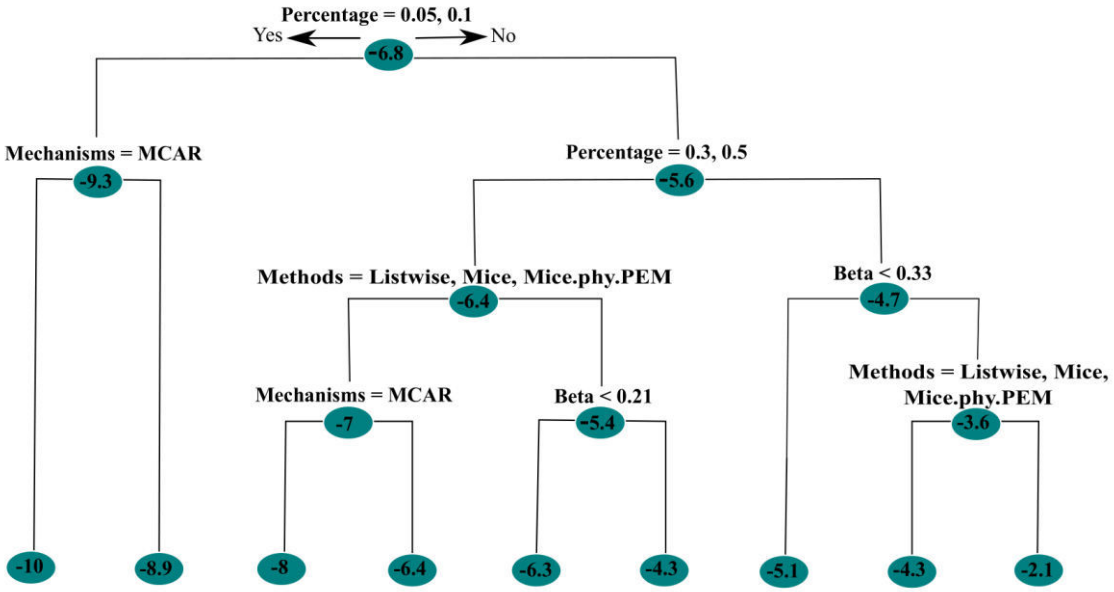


Fig S6. Regression tree analysis of the logarithm of regression coefficient (Beta) squared error (SE). Each node contains the mean of the logarithm SE in each tree split. In the nodes, there are the criterion and the variables used in each split.

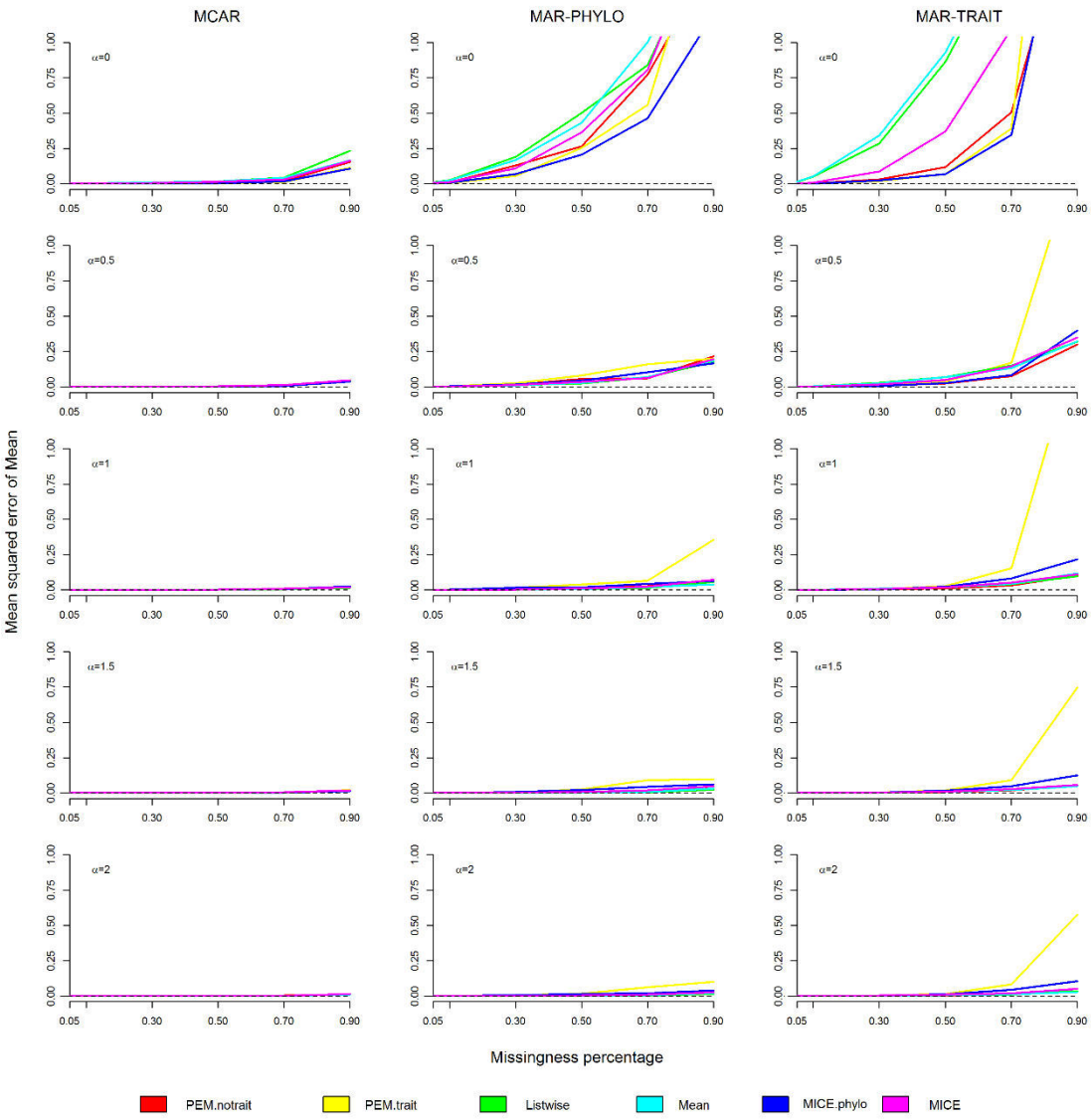


Figure S7. Mean squared error of trait's mean under different methods, OU selective strength, missingness percentage and mechanisms.

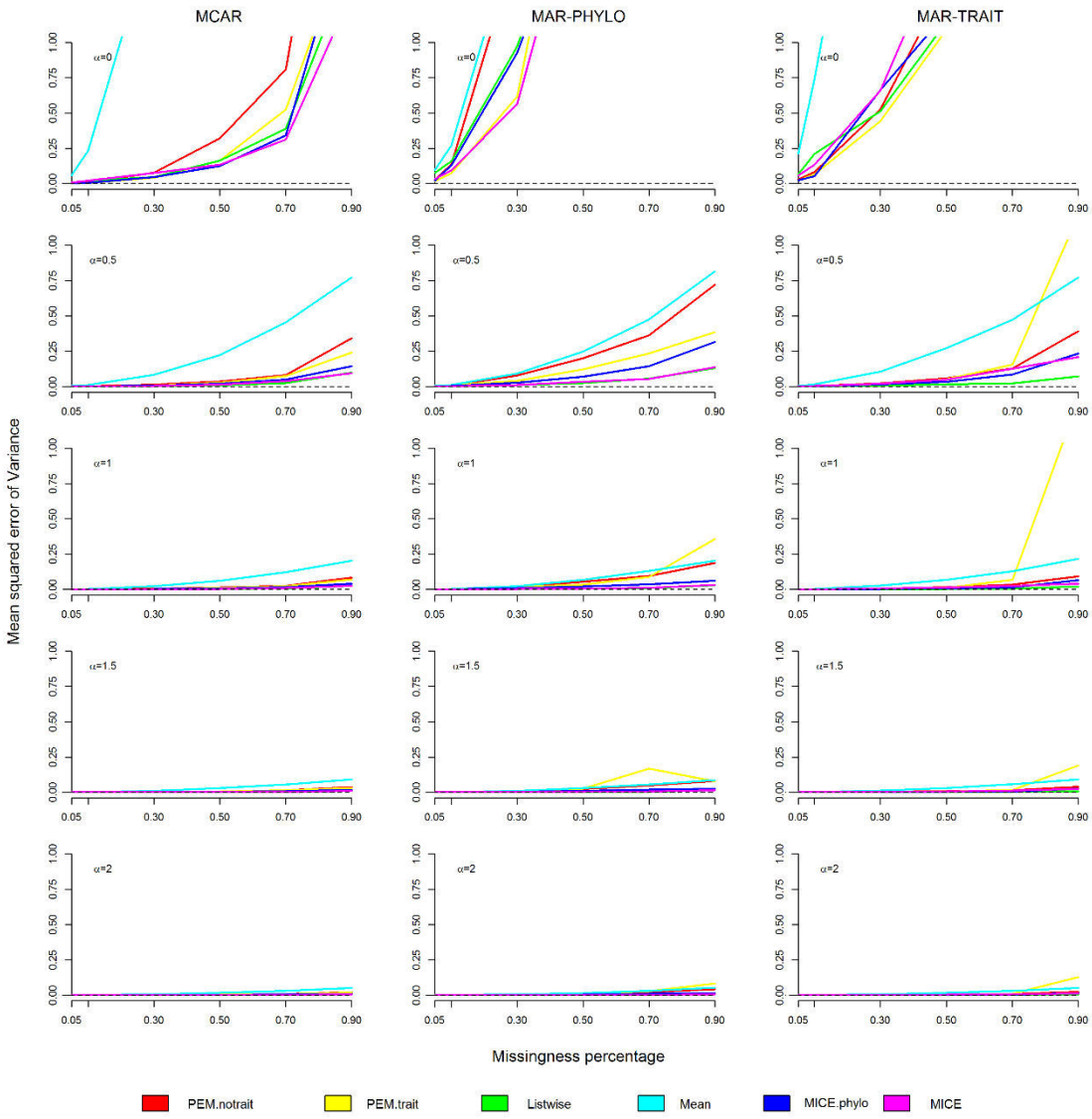


Figure S8. Mean squared error of trait's variance under different methods, OU selective strength, missingness percentage and mechanisms.

APPENDIX 2

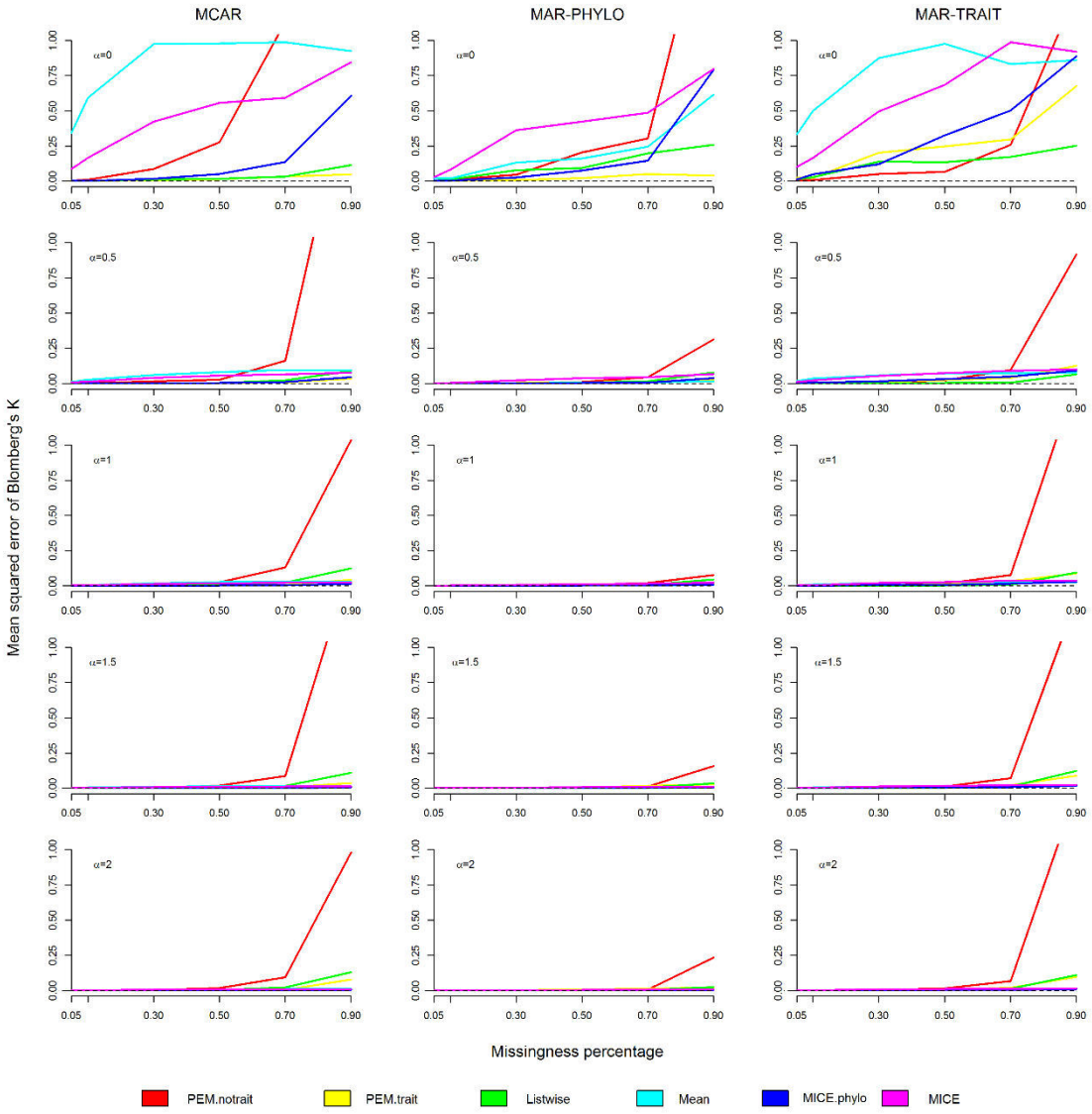


Figure S1. Mean squared error of Blomberg's K under different methods, OU selective strength, missingness percentage and mechanisms.

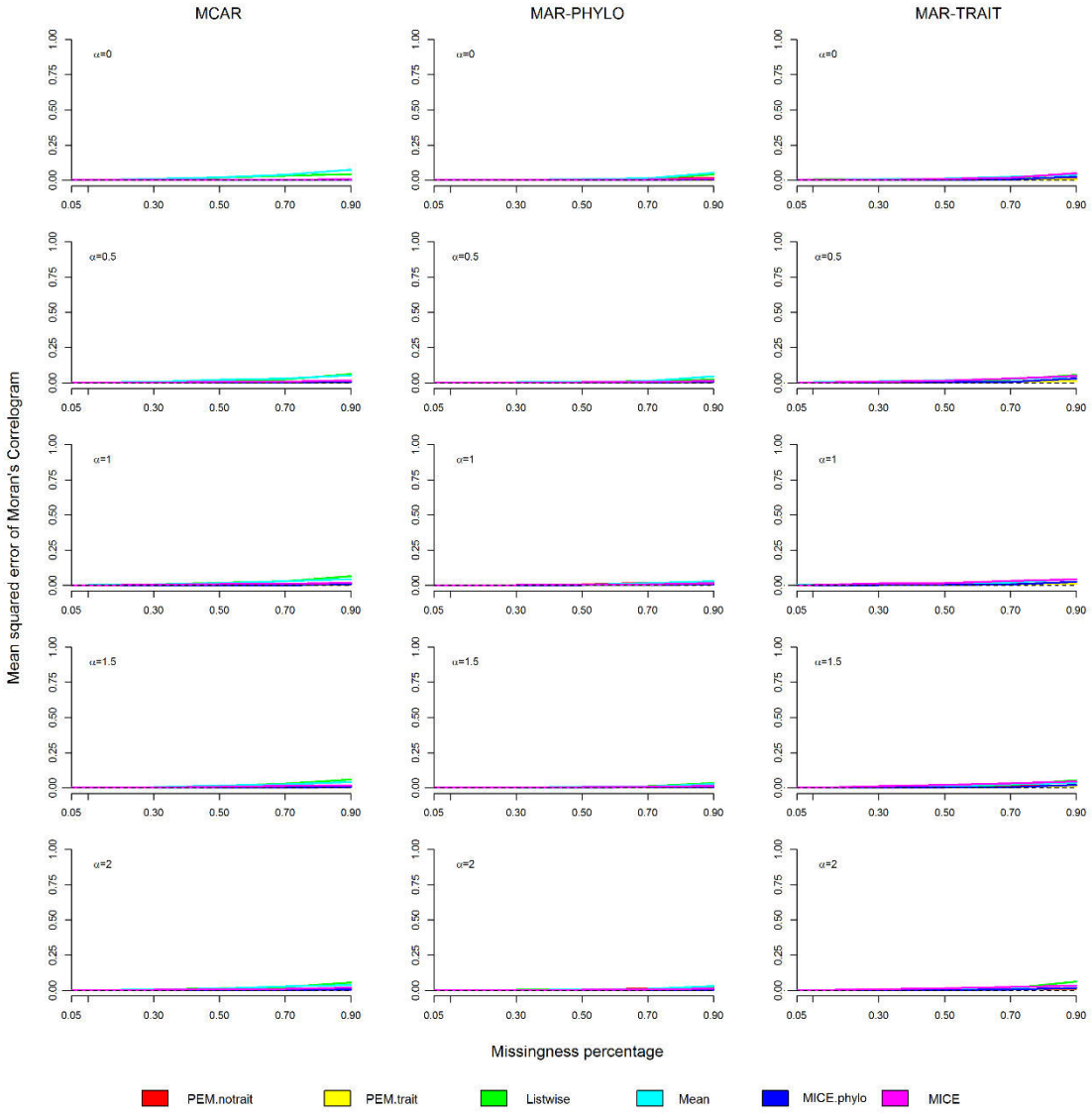


Figure S2. Mean squared error of Moran's Correlogram under different methods, OU selective strength, missing data percentage and mechanisms.

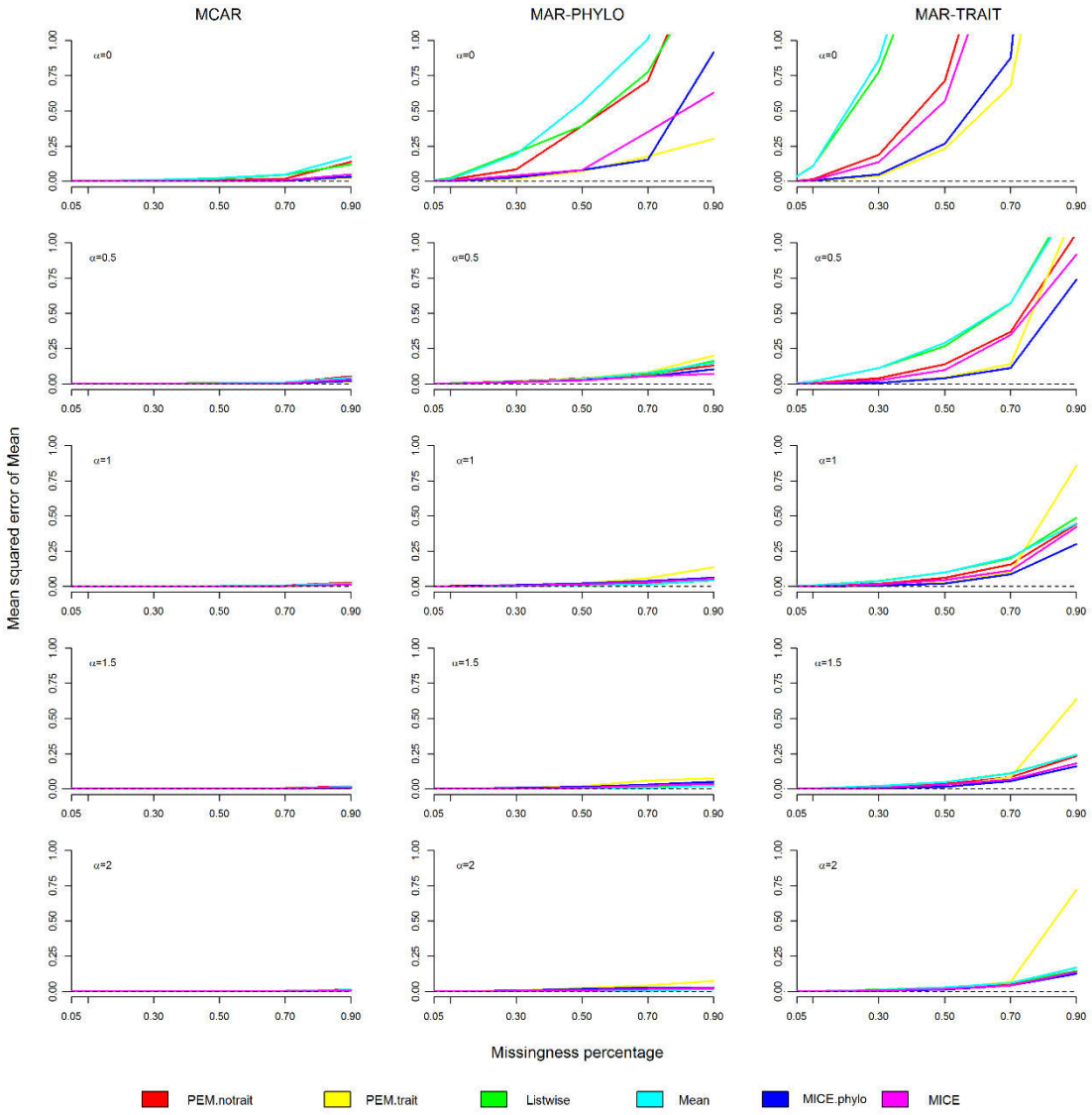


Figure S3. Mean squared error of trait's mean under different methods, OU selective strength, missingness percentage and mechanisms.

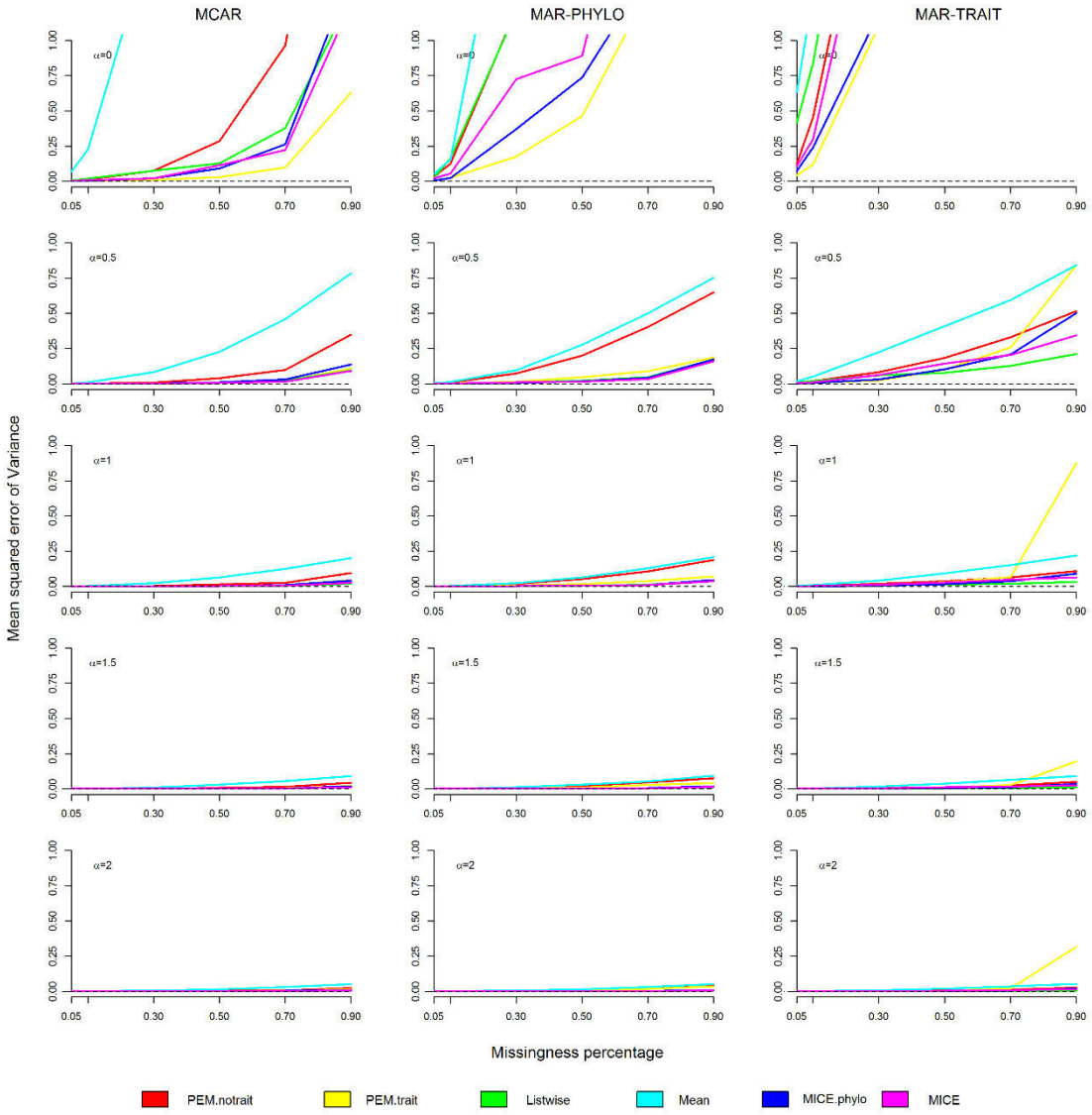


Figure S4. Mean squared error of trait's variance under different methods, OU selective strength, missingness percentage and mechanisms.

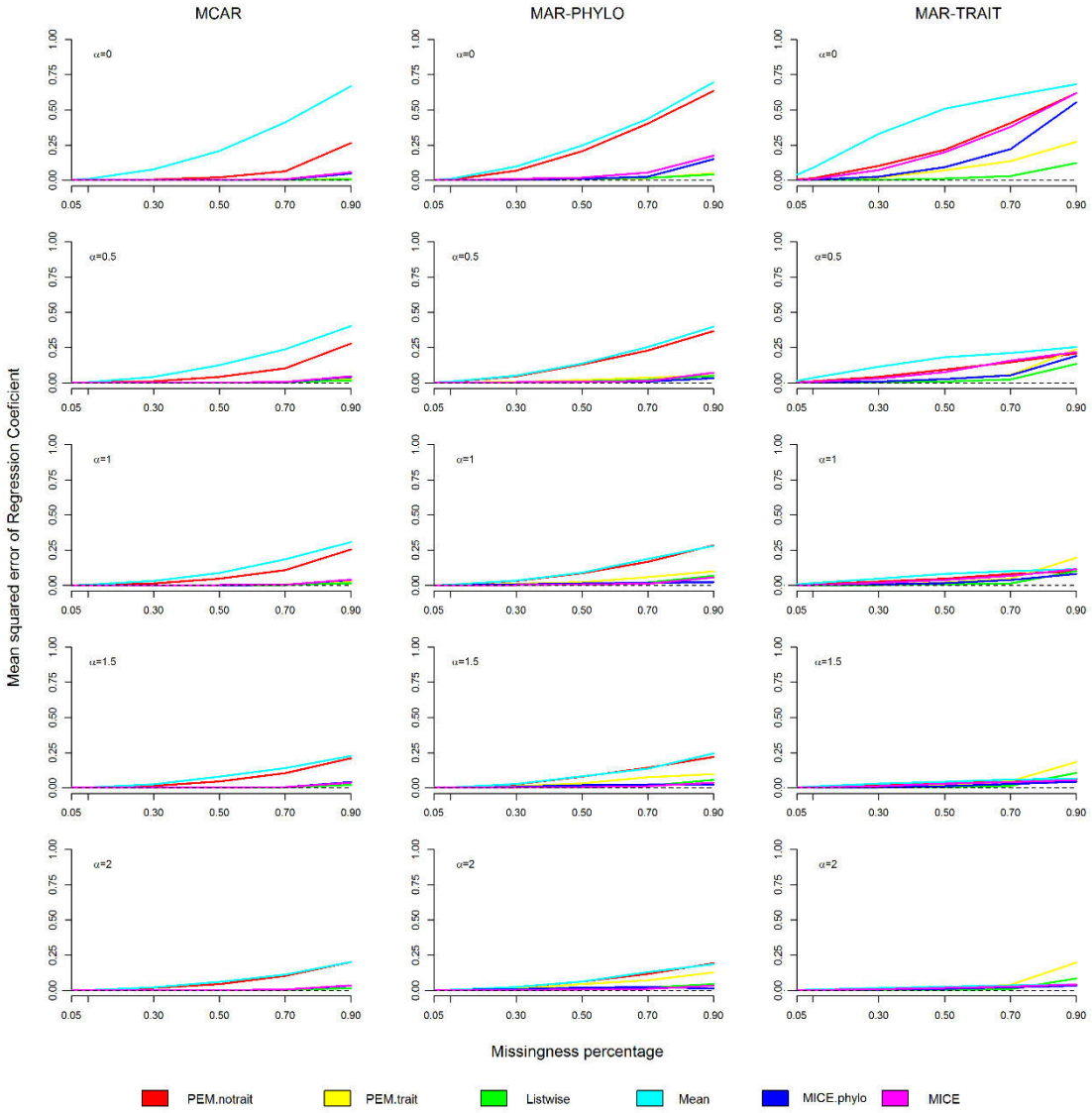


Figure S5. Mean squared error of regression coefficient under different methods, OU selective strength, missingness percentage and mechanisms.

Phylogenetic imputation and brain-body size evolution in primates, with special reference to *Homo floresiensis*

Lucas Jardim[1], José Alexandre Felizola Diniz-Filho[2]

1.Programa de Pós-Graduação em Ecologia & Evolução, Instituto de Ciências Biológicas, Universidade Federal de Goiás.

2.Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás.

Abstract

One of the most famous ecogeographical pattern is Island rule, which states that there is an inverse relationship between body sizes of a species inhabiting mainland and the body size of their relative living in islands. This pattern became more attractive after the discovery of *Homo floresiensis*, a small-bodied hominid from Flores Island, Indonesia. Despite being a probable example of Island rule acting on the genus *Homo*, *Homo floresiensis* has also a smaller brain size than expected by its allometric scaling. Consequently, there has been questions about the appliance of Island rule on Flores man, due to uncertainty about its ancestry and brain-body size relationship. Here we modeled body and brain sizes evolution in primates, including fossil hominids, as a stochastic process model. We included into the models shifts in evolutionary rates across body and brain size evolution detected by Phylogenetic Signal Curve representation (PSR). Yet, we modeled shifts in brain-body relationship and island effects on brain and body sizes and

their evolutionary rates. We selected the best models by Akaike Criterion Information and model adequacy to recover observed patterns. Then, we predicted body and brain sizes of *Homo floresiensis* and compared them to observed values, assuming different hypothesis about its ancestor. Our results showed that primates do not follow Island rule, but early in the *Homo* lineage there have been shifts in evolutionary rates and brain-body relationship. Our models predicted, on average, larger body and brain sizes of *H. floresiensis* than observed values. Thus, Flores man might have been selected toward smaller body and brain sizes in comparison to its mainland ancestor since it colonized Flores Island. Nonetheless, observed brain and body sizes were within predictive confidence intervals of our models, suggesting that *Homo floresiensis* was not an outcome of exceptional evolution.

Keywords: *Homo floresiensis,* Phylogenetic Comparative Methods, Phylogenetic Signal Curve, evolutionary shifts, brain allometry, hominids

Introduction

Island Rule is one of the most famous biogeographical patterns, originally described in the middle 1960's (Foster 1964; Van Valen 1973). According to this pattern, a shift in body size is expected after colonization of islands, eventually leading to the origination of new species better adapted to particular environmental conditions in islands. Lomolino (1985, 2005) proposed that the pattern is actually more continuous than discrete, so that a negative correlation between body size in the mainland and relative body size shift in the island exists. So, large-bodied species will tend to reduce their body size (dwarfing), whereas a small-bodied ancestor will evolve towards a larger new species (gigantism) in islands. The main explanation to dwarfing is that natural selection tends to

favor small-bodied individuals in islands with reduced availability of resources, and because of the high intraspecific competition there is a reduced population abundance and increasing fitness for lower growth rates and lower maturity age (Palkovacs 2003).

The Island Rule tends to be observed in several groups of organisms (Lomolino et al. 2013), but have been most consistently studied for mammals (see Faurby and Svenning (2016) for a recent analysis). The most spectacular evidences of dwarfism come from large mammals such as Proboscideans and several Artiodactyls, as well as from a few clear gigantism cases in Insectivores and Rodents (see van der Geer et al. 2011 for a review). Primates also tend to display Island Rule, with some cases of dwarfing in both body and brain sizes (Bromham and Cardillo 2007; Welch 2009; Montgomery et al. 2010)

The discussions around Island Rule in primates have been amplified by the discovery of a new human species, called *Homo floresiensis* (Brown et al. 2004). This new species was described in 2004 based on a skull (LB-1) and some postcranial material found in a small island in Indonesia, the Flores Island. The skeletal remains, probably from a female, revealed initially a very small-bodied hominid (estimates of about 27 kg, with a brain size of 400 cc), probably derived from a population of *H. erectus* that suffered island dwarfism (Brown et al. 2004). Beyond the small size and the possibility that "island effect" applies to hominids, the discovery was also controversial due to the fact that fossils were considered relatively recent, ca. about 70-90 kya (see Sutikna et al. 2016). However, more recently van den Bergh et al. (2016) found new fragments of *H. floresiensis* that dated for 700 kya, which ended some previous discussion about the validity of the new species, which was attributed by some researchers to a pathological microcephalic form of *Homo sapiens* (Aiello 2010).

Bromham and Cardillo (2007) pointed out that this shift in body size (and brain size as well) in *H. floresiensis* is within the range of reduction in island primates. Kubo

et al. (2013) and Montgomery (2013) showed that the brain size in *H. floresiensis* seems to be smaller than expected by the dwarfism from a *H. erectus* ancestor based on allometric scaling. Indeed, Martin et al. (2006*a*, 2006*b*) had already used this same reasoning to argue against *H. floresiensis* as a valid species. In addition, recent analyses using evolutionary quantitative genetics approach show that the intensity of directional selection driving observed dwarfing is plausible, even considering a wide range of population parameters and colonization scenarios (Diniz-Filho and Raia 2017). Hence, brain size reduction would not be due to allometric effects alone, and direct selective forces would drive brain size evolution towards smaller sizes, in addition to dwarfism driven by body size reduction as independent of body size (Grabowski 2016; Diniz-Filho and Raia 2017). Indeed, if the main explanation for dwarfism under Island Rule is reducing energetic budget, then a strong reduction in brain size is also expected by considering that cerebral tissues are quite demanding in this sense. This pattern in brain size reduction was also observed for other mammals (Weston and Lister 2009), so it is even possible to hypothesize that brain energetic requirement would be the main driver the overall reduction in body size (Herculano-Houzel and Kaas 2011; Grabowski 2016; Diniz-Filho and Raia 2017).

However, new controversies around the species arose. Since first description in 2004, some researchers have actually suggested that *H. floresiensis* is more anatomically related to older African hominids, including some forms of early *Homo* and *H. habilis* (Argue et al. 2009; Morwood and Jungers 2009; Trueman 2010). Recent cladistics analyses based on cranial characters provided conflicting results, either supporting that *H. floresiensis* belongs to the *H. erectus* clade (Zeitoun et al. 2016) or suggesting an older ancestry, closer to basal African early *Homo* (Dembo et al. 2015; Argue et al. 2017). If this last hypothesis proves correctly, the small body of *H. floresiensis* would not be due

to island ecological processes (Palkovacs 2003; Raia and Meiri 2011; Lomolino et al. 2012) and only reflect deeper ancestry followed by stasis. On the other hand, it is important to highlight that supporting this last hypothesis would also have deep implications for the so called "Out of African I" hypothesis, in which *H. erectus* was the first hominin leaving Africa (see Carotenuto et al. (2016) for a recent analysis and review).

Because of the discussions around the evolutionary relationships between *H. floresiensis* and the other hominids, as well as the issues related to brain size dwarfism independent of body size, it is still important to use different approaches to evaluate models of body and brain size evolution in an explicit phylogenetic context. If it is possible to successfully fit models of brain and body size evolution to primate clade, including hominids, it is possible to verify whether observed data from *H. floresiensis* falls within the expected patterns of primate evolution. In a more methodological context, it is possible to use the "phylogenetic imputation" approach (Garland, Jr., and Ives 2000; Guénard et al. 2013; Swenson 2014; Schrodt et al. 2015) to estimate the value for a given taxa of interest and then compare the observed and estimate values. This allows defining if this taxon significantly differs from the phylogenetic expectations for a given trait, so it requires "ad hoc" explanations based on particular selective forces in that lineage (Vining and Nunn 2016)

Thus, our goal here is to perform a comparative phylogenetic analysis of brain and body size evolution in primates (including some fossil hominids), in the context of island rule. We initially evaluated patterns of evolution in these traits by several nested models with increasing level of complexity, starting from a simple Brownian motion model for interspecific variation in body size up to a more complex model of brain size evolution driven by island rule and complex models of non-stationary brain size evolution. After

comparing these models, we used them to predict the expected values of brain and body size in *H. floresiensis* and to evaluate alternative scenarios and hypotheses for its evolution, and compared such values with the observed ones.

Methods

Insularity definition, body and brain size

Primate insularity was defined, following previous studies, as those species currently living on islands ("classical definition"; Faurby and Svenning 2016). This criterion comprised 41 species including those inhabiting large islands, such as Madagascar and Java, which represent worth independent evolutionary lineages such as Strepsihrrini (e.g.: lemur, galagos) as well as insular Hominidea (*Pongo*.sp).

We gathered body size data from PanTHERIA database (Jones et al. 2009) and from a more recent data compilation (Faurby and Svenning 2016). The species *Chiropotes chiropotes*, *Callicebus discolor*, *Pongo abelii*, *Procolobus kirkii* and *Sciurocheirus gabonensis* presented unreliable body size values, so we checked and supplemented the information by Smithsonian National Museum of Natural History (https://collections.nmnh.si.edu/search/mammals) and All the World's Primates database (Rowe and Myers 2012). Extinct hominin brain and body sizes were gathered from Grabowski et al. (2015, 2016).

Brain size data were gathered from Isler et al. (2008) and Gonzalez-Voyer et al. (2016). In Isler dataset, we excluded *Brachyteles sp*. and *Pygathix sp*. brain sizes because they were genera averages, although we included in our analysis the information about *Pygathrix nigripes* brain size present in their dataset. Gonzalez-Voyer dataset contained

brain masses, so we transformed it to endocranial volume (cubic cylinder) by dividing 1.036 from brain mass (the fresh brain tissue density) (Isler et al. 2008).

Furthermore, we included in our analysis brain size of extinct hominin species to increase the representativeness of *Homo floresiensis* evolutionary pathway, gathered from (Grabowski 2016; Grabowski et al. 2016).

Phylogenetic hypothesis

Our phylogenetic hypothesis was an extant primate phylogeny combined to a fossil hominin phylogeny. The extant primate phylogeny was reconstructed by Springer et al. (2012) using a maximum likelihood inference and a molecular supermatrix for 367 species. The hominin phylogeny was grafted from Dembo et al. (2015) consensus, which was based on craniodental characters for 20 fossil species. Then, we binded hominin phylogeny on the most recent common ancestor between *Pan troglodytes* and *Homo sapiens* on extant phylogeny, but as hominin phylogeny root was slightly younger than *Pan-Homo* node, we re-scaled hominin branch lengths to make node ages compatible.

Dembo et al. (2015) and Argue et al. (2017) estimated the phylogenetic position of *Homo floresiensis* as a possible *Homo* ancestral lineage, nonetheless there is not yet in literature a consensus about *Homo floresiensis* phylogenetic position. Hence, we binded this species in three other hypothetical ancestral scenarios: (1) *Homo habilis*, (2) early *Homo* (Argue et al. 2017) and (3) *Homo erectus/ergaster*.

Body and brain size evolutionary models

We modeled primate body and brain size (log-scale) evolution as a stochastic process (Cavalli-Sforza and Edwards 1967; Felsenstein 1973; Martins and Hansen 1997)

that describes trait evolution as a random-walk through branches of a specified phylogeny. Then, trait evolution is drawn from a normal distribution with mean zero and variance $\sigma^2 t$, where t is the branch length in which trait evolved through. At the end of the process, species traits are expected to have the mean close to the initial value (root state) and the covariance between pairs of species is directly proportional to sum of the branch lengths shared by them. Therefore, this process is well represented by a multivariate-normal distribution (MVN) that is compound by mean and covariance structures (Felsenstein 1973; Martins and Hansen 1997). If mean structure is an $n \times 1$ matrix with equal α values representing n species, α is the trait state at phylogeny root. On the other hand, mean structure can be specified as an equation among trait and explanatory variables representing their global relationship (Martins and Hansen 1997). Moreover, covariance structures is expressed in an $n \times n$ matrix **(V)** whose off-diagonal elements are sum of the shared branch lengths between two species since the root until their most recent common ancestor (Felsenstein 1973; Martins and Hansen 1997). Species variances are described in **V** diagonal elements by sum of the branch lengths since the root until the phylogeny tips, therefore if phylogeny is ultrametric, diagonal is constant and all species have the same variance. However, **V** describes traits evolving with an evolutionary rate (σ^2) equal to one, thus to change evolutionary rate, **V** needs to be multiplied by σ^2. Additionally, evolutionary rates can be different along phylogeny branches and these shifts can be modeled by multiplying different σ^2 by its respective branch length before constructing **V** (O'Meara et al. 2006). Furthermore, one can include a parameter λ to control phylogenetic signal by multiplying the off-diagonal elements of **V** by λ (Pagel 1999).

Taking this framework, we started by modelling body size evolution. To do so, we specified a baseline evolutionary model (Model 1) with mean structure equals to root

state and a constant evolutionary rate along all phylogeny branches (i.e., a classical Brownian-motion model). Then, we evaluated the baseline evolutionary model adequacy (see below) and detected evolutionary rate shifts through primate body size evolution that were not captured by our model. Hence, we included those shifts in a new model (Model 2), so adding them improved the evolutionary model adequacy. We selected best baseline model by second-order Akaike Information Criterion (AICc) and model evidence ratio (Burnham and Anderson 2002).

Once we selected the best baseline model that can describe appropriately the most general body size evolutionary pattern, we created other model (Model 3) adding the island effect on mean structure, so that

$$\log(\text{Body size}) = \alpha + \beta X \qquad \text{eq.1}$$

whereas α is the root state and X is a dummy variable specifying if each primate species is insular or not. Furthermore, we created another model in which island species evolve in a different rate than mainland primates (Model 4). Then, clades in which all species were insular had their insular colonization assumed to be occurred at the most recent common ancestor (MRCA) of the clade. Then, every descendant branch of that MRCA were assumed to be evolved at insular rate. If island colonization happened at terminal branches, that branches were also assigned island evolutionary rate. Thus, we ranked baseline models and these models with island effects by AICc, as abovementioned.

Brain size evolution was also modelled as stochastic process, and we created two initial baseline models: (Model 1) a constant rate model and mean structure equal to root state, and; (Model 2) a model with constant rate and the mean structure equal to an allometric relationship between brain and body size:

$$\log(\text{Brain size}) = \log(\alpha) + \beta\log(\text{Body size}) \qquad \text{eq.2}$$

where α is the intercept and β is the allometric coefficient. Then, we selected the best model and checked its adequacy. As we detected evolutionary shifts that were not accounted by our best baseline model, we also included those shifts into the model (model 3). However, brain-body size relationship showed a different slope for Hominin (here *Australopithecus sp.* and *Homo sp.*) and Pongids (*Pan sp.*, *Pongo sp.* and *Gorilla sp.*), so we included those allometric changes in a model 4. We ranked these three models by AICc and the best model was used as our baseline model in which we included island effects on mean structure (model 5) and evolutionary rates (model 6), as previously done for body size evolution.

We estimated all models by maximum likelihood using the function *mle2* from package *bbmle* (Bolker and R Development Core Team 2017). The maximum likelihood searches were realized by L-BFGS-B (Byrd et al. 1995), which allowed us to constrain λ between 0 and 1, and σ^2 upper than 10^{-3}, because variances close to zero took the searches to regions of very low likelihood where the models could not leave there. To explore possible global optimization problems, we ran the searches of each model from five random starting points and used the maximum likelihood search of each model in the subsequent analysis. All analyses used a phylogeny with its height scaled to one (Fuentes-G. et al. 2016).

Model adequacy

While building the models described above we checked their adequacy to reproduce the observed phylogenetic pattern, because model selection does not guarantee that the best model actually fits data properly (i.e., the best model is not necessarily a good model) (Pennell et al. 2015). To access the adequacy of our models, we followed the framework proposed by (Pennell et al. 2015). He proposed that if a phylogenetic

covariance matrix **V** is transformed properly to represent some trait evolution, we could expect a trait simulated by Brownian-motion with mean zero and σ^2 equal to 1 could recover, in average, diagnostic statistics calculated over the original trait. Here, we used squared mean (squared PIC mean) and coefficient of variation (PIC coefficient of variation) of Phylogenetic Independent Contrasts (PIC) (Felsenstein 1985), besides the linear relationship between PIC and their expected variances (PIC-Variance), as suggested by (Pennell et al. 2015). However, PIC is not straightforward to visualize where evolutionary shifts happened in trait evolution. Therefore, we applied the Phylogenetic Signal Representation (PSR) proposed by (Diniz-Filho et al. 2012) to detect where there were evolutionary shifts not represented by our models, what has been showed to be congruent with other more computationally intensive method (Diniz-Filho et al. 2015) (but see Mazel et al. (2016) for other alternative).

PSR procedure extracts eigenvectors from a phylogenetic-distance matrix and applies regressions among a trait on cumulative eigenvectors, in a decreasing order of their eigenvalues, and calculate their determination coefficients (R^2). Under a Brownian motion process, a linear and directly proportional relationship will appear between cumulative eigenvalues and regression R^2. This procedure is repeated for each simulated and observed traits and absolute deviations between successive eigenvectors are calculated. If some observed deviations has less than 5% of probability to be recovered by simulated deviations, an evolutionary shift is detected (Diniz-Filho et al. 2015).

Thus, we scaled the primate phylogeny to height equal to one, as we did while fitting the models, and transformed the phylogeny branch lengths using the estimated λ and σ^2. Then, we simulated 10000 traits by a Brownian-motion process with mean zero and σ^2 equal to one using *rTraitCont* function of the R package *ape* (Paradis et al. 2004). Thereafter, we calculated PIC and their variances by *pic* function of package *ape* (Paradis

et al. 2004) and the diagnostic statistics (see above) for the simulated values and the model residuals. We considered some model adequate if it could reproduce the observed statistics in 5% of a two-tailed test. If the models were inadequate in respect to variance coefficient or PIC-Variance, it represented non-stationarity of σ^2 along the phylogeny. Therefore, we created PSR curves to detect those shifts and then included them into the model.

Predicting Homo floresiensis body and brain size

To estimate how large *Homo floresiensis* body and brain sizes should be according to our models, we followed (Goldberger 1962; Garland, Jr., and Ives 2000) equation:

$$Y_{Homo\ floresiensis} = \beta X_{Homo\ floresiensis} + \mathbf{V}^{T}_{Homo\ floresiensis}\mathbf{V}^{-1}(Y-Y_{est}) \qquad \text{eq.3}$$

where β is the estimated model parameters, X is the design matrix that could be an $n \times$ vector of ones or a $n \times m$ matrix of m explanatory variables (eg. insularity and body size). $\mathbf{V}^{T}$ is the transposed covariance matrix of transformed phylogeny for *Homo floresiensis* without its variance, $\mathbf{V}^{-1}$ is the inverse covariance matrix of the transformed phylogeny without *Homo floresiensis*, $Y - Y_{est}$ is the model residuals. To estimate prediction uncertainty we drawn 10000 values from a normal distribution with mean $Y_{Homo\ floresiensis}$ and variance equal to the transformed terminal branch length of *Homo floresiensis* and calculated the 5% and 95% quantiles. These calculations were repeated for each *Homo floresiensis* ancestral hypotheses. Nonetheless, *Homo erectus* ancestral scenarios assumed a brain size of 600, 991 and 1054 cc to represent brain size variation since forms of *Homo habilis* and Dmanisi's D2700 (early forms) up to late and large-brained Indonesian forms such as Sangiran 12 (Coqueugniot et al. 2004). Body size of *Homo*

erectus did not vary a lot between early and late forms according to (Grabowski et al. 2015, 2016), so we assumed a mean body size of 51 kg for all analyses

Results

Body size evolutionary model

Our baseline model (1) was not adequate to capture PIC coefficient of variation and PIC-Variance (Table 1, Fig. 1). Actually, we detected six evolutionary rate shifts in body size throughout the phylogeny that the model was not reproducing: the first shift occurred during Infraorder Lemuriformes diversification, comprising the genera *Propithecus*, *Avahi*, *Indri*, *Cheirogaleus*, *Microcebus*, *Mirza*, *Lepilemur* and *Phaner*. The second and third shifts occurred in New World primates, during the splitting of the families Atelidae and Cebidae, which also suffered further later shifts in genera *Saguinus*, *Cebus* and *Callithrix*.

Then, we created models 3 and 4 based on model 2. Model selection showed that model 2 was the best model with an AICc weight of 0.67, meaning this model was 2.91 more probable than the second-ranked model, which is the one that included island effects on mean structure (Table 2). Models 1 and 4 were less supported by our data (Table 2).

We then predicted *Homo floresiensis* body size by the four models and the three ancestor scenarios. We did not found prediction differences among models within each ancestor hypotheses, and all models suggest that *Homo floresiensis* as an early *Homo* ancestor would be roundly a 40 kg primate, and as a *Homo habilis* descendent its size would be as large as 32 kg and as a *Homo erectus* descendent it would weight 50 kg. However, models differ regarding to confidence intervals and even a *Homo erectus*

Table 1. Model adequacy of body and brain size evolutionary models. Each value represents the probability of observed statistics be recovered by model simulations. Model with values higher than 0.05 were considered adequate to recover that statistic.

Trait	Model	Squared PIC mean	PIC coefficient of variation	PIC-Variance	PSR
Body size	1	0.96	0	0	0.06
	2	0.93	0.02	0.18	0.08
Brain size	1	0.96	0	0	0.03
	2	0.96	0	0.86	0.2
	3	0.95	0.33	0.5	0.08
	4	0.97	0.08	0.73	0.49

PIC: Phylogenetic Independent Contrasts
PIC-Variance: Linear regression between PIC and its variance
PSR: Phylogenetic Signal Curve representation

ancestral scenario, in which *H. floresiensis* descend form an ancestor with 50 kg, could have a descendent lineage of 27 kg (i.e., the estimated body size for *Homo floresiensis*) in our two best models (Table 3). The other ancestors also included the observed *Homo floresiensis* size in their prediction quantile intervals.

Brain size evolutionary model

Our two baseline models for brain size (models 1 and 2) were not adequate to capture most evolutionary pattern of brain size evolution (Table 2), but the last model was better supported by our data (Table 4). Then, we used model 2 and diagnosed shifts in Platyhrrini evolution, in *Gorilla*, in the base of Hominin, in *Homo* and in *Callithrix* (Fig. 2).

Table 2: Model selection and estimated parameters of body size evolutionary models. λ estimates phylogenetic signal, σ^2 represents shifts in evolutionary rates across phylogeny. Models were selected by delta AICc (dAICc) and AICc weights (Weights).

Models	dAICc	df	Weight	Intercept (se)	Island (se)	λ	σ^2							
							Primates	Strepsihrrini	Atelidae	Cebidae	*Saguinus*	*Cebus*	*Callithrix*	Island
2	0	9	0.67	7.13 (0.28)	-	1	1.68	3.53	0.33	5.58	0.21	0.52	0.62	-
3	2.12	10	0.23	7.14 (0.28)	-0.04 (0.009)	1	1.68	3.55	0.32	5.71	0.21	0.53	0.62	-
4	4.33	9	0.08	7.11 (0.3)	-	1	1.79	-	0.32	5.68	0.21	0.52	0.62	2.03
1	6.85	3	0.02	6.95 (0.26)	-	1	1.62	-	-	-	-	-	-	-

We included in our model 3 shifts in *Gorilla*, *Australopithecus, Homo* and *Callithrix*, which improved our model in respect to the two baseline models. Moreover, this model was adequate for all diagnostic statistics, even without including all detected Platyhrrini shifts.

Table 3. Predictive quantiles of body size for each evolutionary model and ancestor scenarios.

Models	Ancestral scenarios								
	Homo ancestor			*Homo habilis*			*Homo erectus*		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
2	15.98	40.56	103.05	14.85	32.32	70.87	26.90	50.29	93.69
3	15.50	38.85	97.24	14.05	31.27	69.02	25.81	48.66	91.46
4	24.12	40.66	68.01	20.42	32.37	51.56	37.21	50.37	68.19
1	24.65	40.38	67.04	21.45	32.53	49.20	36.06	50.57	70.71

Our best model was model 4 that did not considered island effects on mean structure neither in evolutionary rates. This model was 2.25 more probable to generate our observed data than the second-ranked model and 3.18 more probable than the third – ranked model, which considered island effects respectively on mean structure and evolutionary rates (Table 4). Despite, the second and third ranked models were not far from the best model.

Brain size evolutionary models predicted larger brain sizes than that observed for *Homo floresiensis* (Table 5), except for the three best models in a scenario in which the ancestor had a brain size of 600 cc (closer to *H. habilis* or early *H. erectus*, as Dmanisi). The first and third ranked models predicted similar brain size estimates, whereas the second-ranked model predicted, on average, 13 cc smaller brain sizes due to island effect on mean structure. Besides, the three worst models, which did not considered Hominin

Table 4. Model selection and estimated parameters of brain size evolutionary models. λ estimates phylogenetic signal, σ^2 represents shifts in evolutionary rates across phylogeny. * represents shifts in body-brain relationship. Models were selected by delta AICc (dAICc) and AICc weights (Weights).

Models	dAICc	d.f.	Weight	Intercept (se)	Hominin (se)	Pongid (se)	Island (se)	Body size	Hominin* Body size	Pongid* Body size	λ	σ^2					
												Primates	*Callithrix*	*Gorilla*	Hominin	*Homo*	Island
4	0	12	0.54	-1.46 (0.05)	-4.58 (3.63)	4.44 (1.11)	-	0.61 (0.0007)	0.51 (0.03)	-0.38 (0.008)	1	0.12	8.6	0.001	2.26	0.39	-
5	1.63	13	0.24	-1.44 (0.05)	-4.61 (3.9)	4.51 (1.10)	-0.03 (0.001)	0.61 (0.0006)	0.51 (0.034)	-0.39 (0.008)	1	0.12	7.9	0.001	1.92	0.44	-
6	2.35	13	0.17	-1.48 (0.05)	-4.52 (3.78)	4.52 (1.11)	-	0.61 (0.0006)	0.50 (0.03)	-0.39 (0.008)	1	0.12	7.49	0.001	2.05	0.42	0.11
3	4.38	8	0.06	-1.49 (0.05)	-	-	-	0.61 (0.0006)	-	-	1	0.12	7.87	0.54	2.48	1.85	-
2	93.96	4	0	-1.67 (0.08)	-	-	-	0.64 (0.001)	-	-	1	0.2	-	-	-	-	-
1	266.4	3	0	2.78 (0.17)	-	-	-	-	-	-	1	1.02	-	-	-	-	-

and Pongids differences on allometric relationships, predicted larger brain sizes than observed. Therefore, if *Homo floresiensis* were ancestor of *Homo* its brain size would be expected roughly between 450 and 470 cc, for the three best models. Otherwise, considering *Homo habilis* as ancestor, the brain size would be actually larger (~ 480-490 cc) than that predicted by *Homo* ancestor scenario. A scenario in which *Homo erectus* had 991 cc of brain size predicted brain sizes between 468 and 480 cc. On the other hand, if the *Homo erectus* ancestor had a smaller brain size (600 cc), the prediction would be between 349 and 360 cc. Finally, *Homo erectus* with a brain size of 1059 cc as ancestor predicted a brain size between 488 and 500 cc. In all cases, quantile intervals of brain size predictions were wide and actually encompassed the observed *Homo floresiensis* brain size, except for model 1.

Discussion

Body size evolution and Island Rule in primates

Island rule has been argued as a general pattern acting on several mammal orders, as well as birds and reptiles (Lomolino 2005; Lomolino et al. 2013; Faurby and Svenning 2016). Primates, in turn, has received little attention on this issue that mostly concluded for the rule appliance (Bromham and Cardillo 2007; Meiri et al. 2008; Nowak et al. 2008; Schillaci et al. 2009; Welch 2009; Montgomery et al. 2010, 2016; Lomolino et al. 2013; Montgomery 2013). However, our results showed it is more plausible that primates does not follow island rule in general, because neither their optimal body size nor evolutionary rates have changed on island systems.

The main critics against island rule states that gigantism and dwarfism on islands are taxa specific, time-dependent and contingent on responses to environmental and biotic pressures instead a rule (Meiri et al. 2006, 2008, 2011; Raia et al. 2010; Raia and Meiri 2011). For instance, *Macaca fascicularis* is a primate that does not follow island rule

Table 5. Predictive quantiles of brain size for each evolutionary model and ancestor scenarios.

Models	Ancestor scenarios													
	Homo ancestor			*Homo habilis*			*Homo erectus* (991 cc)			*Homo erectus* (600 cc)			*Homo erectus (1059 cc)*	
	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
4	361.83	473.44	613.12	388.76	493.66	627.97	394.13	483.1	591.06	294.87	361.48	443.61	409.87	501.18	614.49
5	344.13	456.32	606.08	377.18	482.09	614.25	379.21	468.59	577.81	283.49	349.96	430.6	396.99	488.4	602.22
6	389.17	470.81	571.17	418.7	492.92	581.7	407.91	483.71	573.95	304.46	360.86	430.85	421.79	503.14	598.09
3	321.93	551.86	938.47	339.63	542.32	869.67	439.53	650.06	958.38	320.96	469.49	693.1	461.39	673.77	996.71
2	406.27	518.69	654.43	392.86	496.91	630.72	471.31	583.52	727.03	389.06	485.57	603.57	483.14	601.37	748.44
1	476.92	703.22	1037.83	442.29	613.06	853.53	722.76	942.3	1245.74	521.32	679.62	886.18	754.87	987.32	1294.89

even inhabiting island of different sizes in Sunda Shelf Islands archipelago, Southeast Asia (Schillaci et al. 2009). *Procolobus kirkii*, on the other hand, had its size decreased and evolutionary rates accelerated in comparison to continental relatives (Nowak et al. 2008). Moreover, even though Madagascar primates have evolved from the same

common ancestor (Yoder and Yang 2004), they diversified into dwarf species in the Cheirogaleidae family (Masters et al. 2014), while Lemuridae and Indridae families evolved convergent giant species (Masters et al. 2014), which weighted more than 10 kg and achieved about 200 kg in *Archaeoindris fontoynontii* (Fleagle 2013). Therefore, body size evolution of primates on islands may be lineage dependent, likely reflecting differences of species biological characteristics such as diet, life history and trophic level (Lomolino et al. 2012) that are variable within Primates order. Then, antagonistic responses through primate lineages in respect to island pressures may result in an absence of general island tendency.

However, we cannot actually discard an "island effect", once our second-ranked model, instead less supported, was not so far from the first-ranked model and estimated an average trend to dwarfism in insular primates. Although, we should note that the second-ranked model captured only a directional tendency in island primates, not the centering trend proposed by "classical" island rule (Foster 1964; Lomolino 1985), in which large species decrease their sizes while small species get larger. Consequently, a small or absence of island effect could emerge from a centering tendency or could be guided by some extreme and rare dwarfisms of such as 80% of reduction described in (Bromham and Cardillo 2007). We consider our model was able to capture island rule, if it exists, once if primates follow island rule they would, on average, decrease their sizes (Bromham and Cardillo 2007; Lomolino et al. 2013), as detected by our model. However, size reduction on island was not better supported by our data, because its effect did not increased model likelihood in comparison to the model without island effect as much as it would be necessary to overcome its model complexity.

In addition to body size displacement, island rule also suggests that species would evolve faster on islands (Millien 2006, 2011; Evans et al. 2012; Rozzi and

Lomolino 2017). An example in the primates is *Procolobus kirkii,* which had accelerated cranium features evolution after island colonization (Nowak et al. 2008). However, despite our model that described this hypothesis had indeed estimated a faster evolution on islands, this model received little support by our data. Concordantly, Raia and Meiri (2011) did not found support for acceleration of mammal body size evolution on islands. Nonetheless, body size evolution on island has been described as a process in which in early stage evolves fast and then its rate slows down and maintains constant since then (Millien 2006; Rozzi and Lomolino 2017). This accelerating-decelerating dynamic at terminal branches of a phylogeny may behave like a constant evolutionary rate at macroevolutionary scale that resembles mainland rate. Thus, our models may have captured this pattern. Further studies could model an accelerating-decelerating evolution (Blomberg et al. 2003; Harmon et al. 2010) on island species and see whether it is indistinguishable or better supported when compared to a constant-rate model.

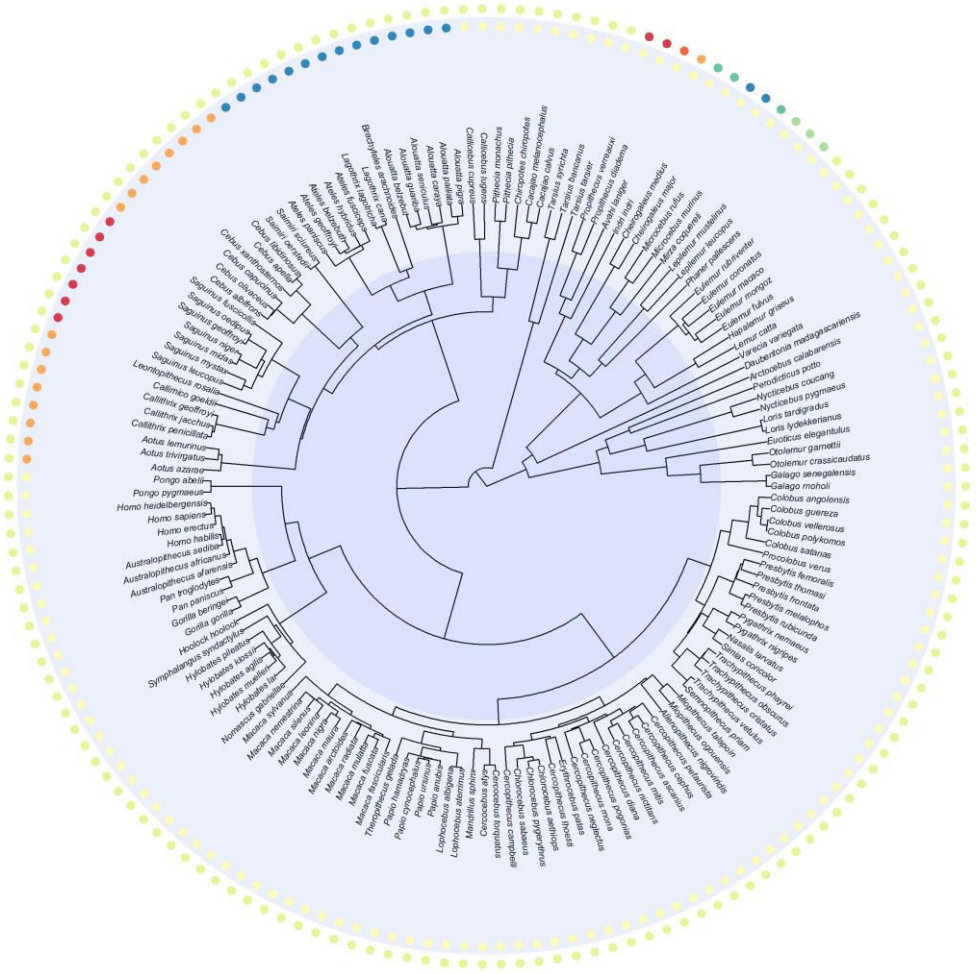


Figure 1. Evolutionary shifts of body size through primate phylogeny. Each line of points represents eigenvectors that increase R^2 of body more than expected by model 1. Red and blue colors mean respectively positive and negative scores. Thus, evolutionary shifts occurred within Infraorder Lemuriformes, family Atelidae, Cebidae and within *Saguinus*. After control for these shifts, there were also detected shifts within *Cebus* and *Callithrix*.

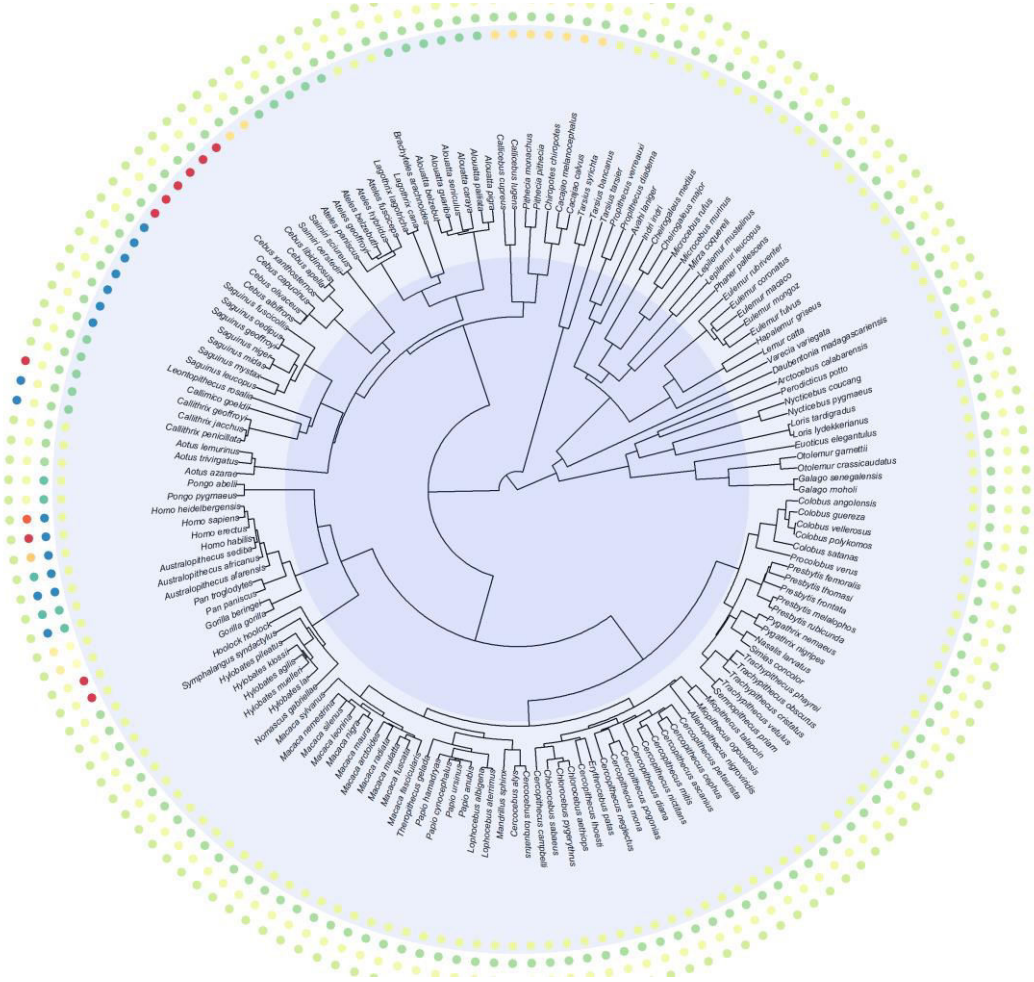


Figure 2. Evolutionary shifts of relative brain size through primate phylogeny. Each line of points represents eigenvectors that increase R^2 of relative brain size more than expected by model 2. Red and blue colors mean respectively positive and negative scores. Thus, evolutionary shifts were detected for Platyrrini, *Callithrix*, Hominin, *Gorilla* and *Homo*.

The most typical (or at least more polemic and widely discussed) example of island rule application on primates is *Homo floresiensis* (Brown et al. 2004; Bromham and Cardillo 2007; Diniz-Filho and Raia 2017) that had its body mass limits estimated ranging from 16 to about 40 kg (Brown et al. 2004; Kubo et al. 2013). Our models predicted masses closer to literature upper limit (~30-50 kg), but these estimates were

extremely dependent on which ancestor *Homo floresiensis* had evolved from, as our models assumed an expected body mass similar to the ancestor. Therefore, evolutionary scenarios whereby *Homo floresiensis* was ancestor of *Homo* or descendant from *Homo habilis* predicted masses within literature ranges, whereas *Homo habilis* ancestor scenario predicted the most accurate mass if *Homo floresiensis* weighted about 27 kg (Grabowski et al. 2015). On the other hand, if *Homo floresiensis* evolved from *Homo erectus* its size would be near to 50 kg, a much larger mass than the literature upper limit, and neither the island effect that we estimated could decrease body mass lesser than 40 kg. Consequently, these results could initially suggest that *Homo floresiensis* was not a dwarf hominin descent from *Homo erectus*, but actually a more ancient hominin as stated by (Dembo et al. 2015; Argue et al. 2017). Nonetheless, our models had predictive quantile intervals that covered literature limits of *Homo floresiensis*, even in scenarios whose ancestor was *Homo erectus* and islands had no effect on body mass evolution. Accordingly, a primate as large as *Homo erectus* are likely to speciate into a *Homo floresiensis*-kind primate without an exceptional, as termed by (Vining and Nunn 2016), body mass evolution. Other studies, that concluded for island rule, have also found that *Homo floresiensis* size could be achieved by a regular island rule dwarfism (Bromham and Cardillo 2007; Montgomery 2013; Diniz-Filho and Raia 2017). Indeed, it is not necessary a strong directional selection on large individuals to decrease *Homo floresiesis* to its estimated size, it is required only a small (or at least plausible) selection strength and time to evolve (Diniz-Filho and Raia 2017).

Based on our models, it is difficult to propose an alternative explanation of how *Homo floresiensis* became smaller without claiming for island pressures. A possible explanation is that Flores pressures were not idiosyncratic in comparison to pressures suffered by primates through evolutionary history. Thus, body changes such as that seen

in *Homo floresiensis* may have already happened in other clades and time periods even in mainland environments (Montgomery et al. 2010). Alternatively, *H. floresiensis* could be exceptional examples of island rule in primates given its body size and diet, which mainly determines island rule effects ((Lomolino et al. 2012).

Brain size evolution

An unfinished debate about *Homo floresiensis* evolution concerns on its small brain size (~426 cc) (Aiello 2010; Kubo et al. 2013; Diniz-Filho and Raia 2017), more specifically on how large it should be under an allometric scaling with body mass. We found that brain-body relationship might not be constant through primate evolution, whereas shifts in Hominin and Pongid clades notoriously improved our model. Brain-body size scaling has been theorized as an outcome of energy income and brain-body costs that are balanced by reducing high demanding organs such as brain and gut, as stated by expensive-tissue hypothesis (Herculano-Houzel and Kaas 2011). Therefore, expensive-tissue hypothesis posits Pongid had their bodies increased to support larger digestive system in order to metabolize more low-energetic food (Aiello and Wheeler 1995; Herculano-Houzel and Kaas 2011). Hominin, on the other hand, had their energetic budget invested to increase brain size rather than gut. However, our results showed that instead investment on body size, shifts on brain size evolutionary rate have guided the allometric scaling shift. While body size had maintained constant evolutionary rates in Pongid and Hominin, brain size had faster evolution in Hominin compared to other Catahrrini, but *Gorilla* had its brain evolution deaccelerated to almost zero, suggesting a strong conservatism on brain size in this taxa. Therefore, *Gorilla*, Australopithecids and *Homo* differences on encephalization quotient resulted from relative brain size evolutionary shifts rather than being a by-product of body size evolution. This hypothesis

is in accordance to (Herculano-Houzel 2012; Grabowski 2016) that proposed that body size evolution was carried by brain size along hominin diversification.

Given the ability to model, in a phylogenetic context, the complex patterns of brain-body correlation, our results about brain size were concordant to those from models in body size evolution. They suggested that primates do not seems to suffer selection to move their ancestral brain size to a new optimum on islands, then island rule would not be applicable to brain size. Our second and third ranked models were close to the first one, but their AICc distances were merely consequence of their small island rule effects, what made them resemble the model without island effect. Furthermore, Montgomery (2013) also found that island rule could not be applicable to brain size in primates (and this is the only study that we have knowledge on island rule effects on primate brain size).

Brain size shrinking in island species has been detected in hippos that suffered intense dwarfism (Weston and Lister 2009), thus due to their diet based on low-energetic income, selection acted to decrease brain size. Diet has been considered a worth factor to determine primates brain size (DeCasien et al. 2017), as primates has a more diversified diet than ungulate, the island effect could not arise as a general trend. Although, it is plausible that some folivorous/frugivorous primates reduce their brains when inhabiting an environment with scarce food, but this does not apply to all primates, whose diet ranges from insectivorous to carnivorous. Nonetheless, one question remains, why the most emblematic and well-studied example of brain reduction in primates is a generalist species (Morwood et al. 2004; Brown and Maeda 2009), *Homo floresiensis*?

Our models predicted larger brain sizes to *Homo floresiensis* than literature upper limit (~426 cc) (Kubo et al. 2013), but a possible ancestrally from 600 cc *Homo erectus* (Dmanisi) predicted a small brain size than the lower limit (~380 cc)(Brown et al. 2004). Previous studies argued that LB1 *Homo floresiensis* brain size was much smaller than

predicted by brain-body allometric scaling of modern humans, ungulates, proboscides (Martin et al. 2006*b*). Kubo et al. (2013), on the other hand, found that it is possible a brain size as large as 426 cc given *Homo floresiensis* body size, in accordance to other studies (Montgomery et al. 2010; Montgomery 2013; Diniz-Filho and Raia 2017). However, this discussion concerns on uncertainty about brain-body scaling and from which ancestor *Homo floresiensis* had diversified. We found that *Homo* clade had a stepper brain-body scaling than primates in general, as abovementioned, thus when size is reduced, brain size indeed decrease more than expected by other primates or mammals. Furthermore, as we found for body size, predictive quantile intervals for brain size, independently of the ancestral or island effect, covered the observed estimates of LB1. Therefore, a 27 kg primate with a brain size as large as 426 cc is within a plausible macroevolutionary scenario derived from brain-body scaling of primates, taking into account evolutionary singularities within Hominin and *Homo* evolution.

Conclusion

Island rule applies for several taxa, mainly mammal species (Lomolino 2005). Also, body size reduction is correlated to brain size dwarfism (Weston and Lister 2009). Here, we showed that, in general, primate body and brain sizes did not follow Island Rule, and that only evolutionary shifts, independent of insularity, are enough to explain the observed distribution of body and brain size in this group. These conclusions have direct impact on the most famous example of primate Island Rule, *Homo floresiensis*. Our models predicted larger body and brain sizes, on average, than observed values, what could indicate Flores man suffered directional evolutionary pressures in Flores Island, a plausible hypothesis as island pressures are more prone to act on larger species (Lomolino 2005; Lomolino et al. 2012), such as *Homo* species. However, *Homo floresiensis* body

and brain sizes are within predictive confidence intervals of our evolutionary models, so even if Flores man suffered Island Rule, it did not have an exceptional, or extremely different evolution than expected by its *Homo* relatives.

Further studies could test evolutionary rate acceleration and deceleration within primate macroevolution in island and evaluate island effect continuously rather than using a discrete form. Furthermore, studies could look for evolutionary shift within *Homo floresiensis* own lineage and compare it with quantitative genetic model already published, such as Diniz-Filho and Raia (2017). Therefore, we did not solve the mystery about whether *Homo floresiensis* follows Island Rule, but we shed more light on the complex primate evolution and how it can explain in the future the idiosyncrasy of Flores man.

Acknowledgements

L.J. is supported by a CAPES Doctoral fellowship. JAFD.-F has been continuously supported by CNPq Productivity Grants. This work was developed in the context of National Institutes for Science and Technology (INCT) in Ecology, Evolution and Biodiversity Conservation, supported by MCTIC/CNpq (proc. 465610/2014-5).

Literature cited

Aiello, L. C. 2010. Five years of *Homo floresiensis*. American Journal of Physical Anthropology 142:167–179.

Aiello, L. C., and P. Wheeler. 1995. The expensive-tissue hypothesis: the brain and the digestive system in human and primate evolution. Current Anthropology 36:199–221.

Argue, D., C. P. Groves, M. S. Y. Lee, and W. L. Jungers. 2017. The affinities of *Homo floresiensis* based on phylogenetic analyses of cranial, dental, and postcranial characters.

Journal of Human Evolution 107:107–133.

Argue, D., M. J. Morwood, T. Sutikna, and E. W. Saptomo. 2009. *Homo floresiensis* : a cladistic analysis. Journal of Human Evolution 57:623–639.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Bolker, B., and R Development Core Team. 2017. bbmle: tools for general maximum likelihood estimation. R package.

Bromham, L., and M. Cardillo. 2007. Primates follow the " island rule ": implications for interpreting Homo floresiensis 398–400.

Brown, P., and T. Maeda. 2009. Liang Bua *Homo floresiensis* mandibles and mandibular teeth: a contribution to the comparative morphology of a new hominin species. Journal of Human Evolution 57:571–596.

Brown, P., T. Sutikna, M. J. Morwood, R. P. Soejono, Jatmiko, E. Wayhu Saptomo, and R. Awe Due. 2004. A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. Nature 431:1055–1061.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference (2nd ed.). Springer-Verlag, New York, NY.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu. 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing 16:1190–1208.

Carotenuto, F., N. Tsikaridze, L. Rook, D. Lordkipanidze, L. Longo, S. Condemi, and P. Raia. 2016. Venturing out safely: The biogeography of *Homo erectus* dispersal out of Africa. Journal of Human Evolution 95:1–12.

Cavalli-Sforza, L. L., and W. F. Edwards. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570.

Coqueugniot, H., J.-J. Hublin, F. Veillon, F. Houët, and T. Jacob. 2004. Early brain growth in *Homo erectus* and implications for cognitive ability. Nature 431:299–302.

DeCasien, A. R., S. A. Williams, and J. P. Higham. 2017. Primate brain size is predicted by diet but not sociality. Nature Ecology & Evolution 1:112.

Dembo, M., N. J. Matzke, A. O. Mooers, and M. Collard. 2015. Bayesian analysis of a

morphological supermatrix sheds light on controversial fossil hominin relationships. Proceedings of the Royal Society B 282:20150943.

Diniz-Filho, J. A. F., D. M. C. C. Alves, F. Villalobos, M. Sakamoto, S. L. Brusatte, and L. M. Bini. 2015. Phylogenetic eigenvectors and non-stationarity in the evolution of theropod dinosaur skulls. Journal of Evolutionary Biology 28:1410–1416.

Diniz-Filho, J. A. F., and P. Raia. 2017. Island Rule, quantitative genetics and brain–body size evolution in *Homo floresiensis*. Proceedings of the Royal Society B: Biological Sciences 284:20171065.

Diniz-Filho, J. A. F., T. F. Rangel, T. Santos, and L. M. Bini. 2012. Exploring pattern of interspecific variation in quantitative traits using sequential phylogenetic eigenvector regressions. Evolution 66:1079–1090.

Evans, A. R., D. Jones, A. G. Boyer, J. H. Brown, D. P. Costa, S. K. M. Ernest, E. M. G. Fitzgerald, et al. 2012. The maximum rate of mammal evolution. Proceedings of the National Academy of Sciences 109:4187–4190.

Faurby, S., and J. C. Svenning. 2016. Resurrection of the island rule: human-driven extinctions have obscured a basic evolutionary pattern. The American Naturalist 187:812–820.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics 25:471–492.

———. 1985. Phylogenies and the comparative method. The American Naturalist 125:1–15.

Fleagle, J. G. 2013. Primate adaptation and evolution (3rd ed.). Academic Press.

Foster, J. B. 1964. Evolution of mammals on islands. Nature 202:234–235.

Fuentes-G., J. A., E. A. Housworth, A. Weber, and E. P. Martins. 2016. Phylogenetic ANCOVA: estimating changes in evolutionary rates as well as relationships between traits. The American Naturalist 188:615–627.

Garland, Jr., T., and A. R. Ives. 2000. Using the past to predict the present: Confidence Intervals for regression equations in phylogenetic comparative methods. The American Naturalist 155:346–364.

Goldberger, A. S. 1962. Best linear unbiased prediction in the generalized linear regression model. Journal of the American Statistical Association 57:369–375.

Gonzalez-Voyer, A., M. González-Suárez, C. Vilà, and E. Revilla. 2016. Larger brain size indirectly increases vulnerability to extinction in mammals. Evolution 70:1364–1375.

Grabowski, M. 2016. Bigger brains led to bigger bodies?: the correlated evolution of human brain and body size. Current Anthropology 57:000–000.

Grabowski, M., K. G. Hatala, W. L. Jungers, and B. G. Richmond. 2015. Body mass estimates of hominin fossils and the evolution of human body size. Journal of Human Evolution 85:75–93.

Grabowski, M., K. L. Voje, and T. F. Hansen. 2016. Evolutionary modeling and correcting for observation error support a 3/5 brain-body allometry for primates. Journal of Human Evolution 94:106–116.

Guénard, G., P. Legendre, and P. Peres-Neto. 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. Methods in Ecology and Evolution 4:1120–1131.

Harmon, L. J., J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, K. H. Kozak, et al. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Herculano-Houzel, S. 2012. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. Proceedings of the National Academy of Sciences 109:10661–10668.

Herculano-Houzel, S., and J. H. Kaas. 2011. Gorilla and orangutan brains conform to the primate cellular scaling rules: Implications for human evolution. Brain, Behavior and Evolution 77:33–44.

Isler, K., E. Christopher Kirk, J. M. A. Miller, G. A. Albrecht, B. R. Gelvin, and R. D. Martin. 2008. Endocranial volumes of primate species: scaling analyses using a comprehensive and reliable data set. Journal of Human Evolution 55:967–978.

Jones, K. E., J. Bielby, M. Cardillo, S. a. Fritz, J. O'Dell, C. D. L. Orme, K. Safi, et al. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of

extant and recently extinct mammals. Ecology 90:2648–2648.

Kubo, D., R. T. Kono, and Y. Kaifu. 2013. Brain size of *Homo floresiensis* and its evolutionary implications. Proceedings of the Royal Society B: Biological Sciences 280:20130338–20130338.

Lomolino, M. V. 1985. Body size of mammals on islands: The Island Rule reexamined. The American Naturalist 125:310–316.

———. 2005. Body size evolution in insular vertebrates: Generality of the Island rule. Journal of Biogeography 32:1683–1699.

Lomolino, M. V, A. A. Van Der Geer, G. A. Lyras, M. R. Palombo, D. F. Sax, and R. Rozzi. 2013. Of mice and mammoths : generality and antiquity of the island rule. Journal of B 40:1427–1439.

Lomolino, M. V, D. F. Sax, M. R. Palombo, and A. A. Van Der Geer. 2012. Of mice and mammoths : evaluations of causal explanations for body size evolution in insular mammals. Journal of Biogeography 39:842–854.

Martin, R. D., A. M. MacLarnon, J. L. Phillips, and W. B. Dobyns. 2006. Flores hominid: New species or microcephalic dwarf? The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology 288A:1123–1145.

Martin, R. D., A. M. Maclarnon, J. L. Phillips, L. Dussubieux, P. R. Williams, and W. B. Dobyns. 2006*b*. Comment on "The Brain of LB1, *Homo floresiensis*." Science 312:999; author reply 999.

Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interespecific data. The American Naturalist 149:646–667.

Masters, J. C., F. Génin, D. Silvestro, A. M. Lister, and M. DelPero. 2014. The red island and the seven dwarfs: Body size reduction in Cheirogaleidae. Journal of Biogeography 41:1833–1847.

Mazel, F., T. J. Davies, D. Georges, S. Lavergne, W. Thuiller, and P. R. Peres-Neto. 2016. Improving phylogenetic regression under complex evolutionary models. Ecology 97:286–293.

Meiri, S., N. Cooper, and A. Purvis. 2008. The island rule : made to be broken ?

Meiri, S., T. Dayan, and D. Simberloff. 2006. The generality of the island rule reexamined. Journal of Biogeography 33:1571–1577.

Meiri, S., P. Raia, and A. B. Phillimore. 2011. Slaying dragons : limited evidence for unusual body size evolution on islands. Journal of Biogeography 38:89–100.

Millien, V. 2006. Morphological evolution is accelerated among island mammals. PLoS Biology 4:1863–1868.

———. 2011. Mammals evolve faster on smaller islands. Evolution 65:1935–1944.

Montgomery, S. H. 2013. Primate brains, the "island rule" and the evolution of *Homo floresiensis*. Journal of Human Evolution 65:750–760.

Montgomery, S. H., I. Capellini, R. a Barton, and N. I. Mundy. 2010. Reconstructing the ups and downs of primate brain evolution: implications for adaptive hypotheses and *Homo floresiensis*. BMC biology 8:9.

Montgomery, S. H., N. I. Mundy, and R. A. Barton. 2016. Brain evolution and development : adaptation , allometry and constraint.

Morwood, M. J., and W. L. Jungers. 2009. Conclusions : implications of the Liang Bua excavations for hominin evolution and biogeography. Journal of Human Evolution 57:640–648.

Morwood, M. J., R. P. Soejono, R. G. Roberts, T. Sutikna, C. S. M. Turney, K. E. Westaway, W. J. Rink, et al. 2004. Archaeology and age of a new hominin from Flores in eastern Indonesia. Nature 431:1087–1091.

Nowak, K., A. Cardini, and S. Elton. 2008. Evolutionary acceleration and divergence in *Procolobus kirkii*. International Journal of Primatology 29:1313–1339.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Palkovacs, E. P. 2003. Explaining adaptive shifts in body size on islands : a life history approach. Oikos 103:37–44.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Pennell, M. W., R. G. FitzJohn, W. K. Cornwell, and L. J. Harmon. 2015. Model adequacy and the macroevolution of angiosperm functional traits. The American Naturalist 186:E33–E50.

Raia, P., F. Carotenuto, and S. Meiri. 2010. One size does not fit all : no evidence for an optimal body size on islands. Global Ecology and Biogeography 19:475–484.

Raia, P., and S. Meiri. 2011. The tempo and mode of evolution: body sizes of island mammals. Evolution 65:1927–1934.

Rowe, N., and M. Myers. 2012. All the world's primates. Primate Conservation Inc., Charlestown RI. www.alltheworldsprimates.org.

Rozzi, R., and M. V Lomolino. 2017. Rapid dwarfing of an insular mammal – The feral cattle of Amsterdam Island. Scientific Reports 1–8.

Schillaci, M. A., E. Meijaard, and T. Clark. 2009. The effect of island area on body size in a primate species from the Sunda Shelf Islands. Journal of Biogeography 36:362–371.

Schrodt, F., J. Kattge, H. Shan, F. Fazayeli, J. Joswig, A. Banerjee, M. Reichstein, et al. 2015. BHPMF - a hierarchical bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. Global Ecology and Biogeography 24:1510–1521.

Springer, M. S., R. W. Meredith, J. Gatesy, C. A. Emerling, J. Park, D. L. Rabosky, T. Stadler, et al. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS ONE 7:e49521.

Sutikna, T., M. W. Tocheri, M. J. Morwood, E. W. Saptomo, Jatmiko, R. D. Awe, S. Wasisto, et al. 2016. Revised stratigraphy and chronology for *Homo floresiensis* at Liang Bua in Indonesia. Nature 532:366–369.

Swenson, N. G. 2014. Phylogenetic imputation of plant functional trait databases. Ecography 37:105–110.

Trueman, J. W. H. 2010. A new cladistic analysis of *Homo floresiensis*. Journal of Human Evolution 59:223–226.

van den Bergh, G. D., Y. Kaifu, I. Kurniawan, R. T. Kono, A. Brumm, E. Setiyabudi, F. Aziz, et al. 2016. *Homo floresiensis*-like fossils from the early Middle Pleistocene of Flores. Nature 534:245–248.

van der Geer, A., G. Lyras, J. de Vos, and M. Dermitzakis. 2011. Evolution of island mammals: Adaptation and extinction of pacental mammals on islands (1st ed.). Wiley-Blackwell, Chichester, UK.

Van Valen, L. 1973. Pattern and the balance of nature. Evolutionary Theory 1:31–49.

Vining, A. Q., and C. L. Nunn. 2016. Evolutionary change in physiological phenotypes along the human lineage. Evolution, Medicine, and Public Health 2016:312–324.

Welch, J. J. 2009. Testing the island rule: primates as a case study. Proceedings. Biological sciences / The Royal Society 276:675–682.

Weston, E. M., and A. M. Lister. 2009. Insular dwarfism in hippos and a model for brain size reduction in Homo floresiensis. Nature 459:85–88.

Yoder, A. D., and Z. Yang. 2004. Divergence dates for Malagasy lemurs estimated from multiple gene loci: Geological and evolutionary context. Molecular Ecology 13:757–773.

Zeitoun, V., V. Barriel, and H. Widianto. 2016. Phylogenetic analysis of the calvaria of *Homo floresiensis*. Comptes rendus - Palevol 15:555–568.

Cross-species evaluation of Bergmann's rule in mammals: looking at biodiversity knowledge shortfall implications

Lucas L. C. Z. Jardim[1], Fabrício Villalobos[3], José Alexandre F. Diniz-Filho[2]

1. Programa de Pós-Graduação em Ecologia & Evolução, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Departamento de Ecologia. Goiânia, GO, Brasil.

2. Universidade Federal de Goiás, Instituto de Ciências Biológicas, Departamento de Ecologia. Goiânia, GO, Brasil.

3. Red de Biología Evolutiva, Instituto de Ecología, A.C., Carretera antigua a Coatepec 351, El Haya, 91070 Xalapa, Veracruz, Mexico

*Corresponding author: lucas.ljardim9@gmail.com (Lucas L. C. Z. Jardim)

ABSTRACT

Climate forces species to adapt their niches to climatic changes. Karl Bergmann propose that larger species would be selected in cold temperature regions, while small species would be selected in warmer climates, a pattern latter named Bergmann's rule. Here we tested Karl Bergmann hypothesis using phylogenetic generalized linear squares, taking into account the possible consequences of ignoring missing values. We considered missing values using Multiple Imputation by Chained Equations, which predicted missing values using life-history traits and phylogenetic eigenvectors as predictors. Our results did not support Bergmann' rule, independent of regarding missing values. However, missing values uncertainty affected about 50 % of temperature effect on body size. Furthermore, we found ignoring missing values might bias temperature-body size relationship. Therefore, macroecological and macroevolutionary researches would be

improved by regarding missing values, thus methodological and theoretical researches on phylogenetic multiple imputation are a fruitful research field to advance.

Keywords: Bergmann's rule, mammals, body size evolution, Multiple Imputation by Chained Equation, Phylogenetic Comparative Methods, Phylogenetic Generalized Least Squares, Phylogenetic Eigenvectors

INTRODUCTION

Climate imposes selective pressures on species ecological niches through time (Cooper *et al.*, 2011; Araújo *et al.*, 2013), which may force species' physiology, behavior and morphology (Porter & Kearney, 2009; Huey *et al.*, 2012) to adapt to climatic change or, as stated by "Court Jester hypothesis" (Barnosky, 2001; Benton, 2009), they become extinct. Following the reasoning of the climate pressure, Karl Bergmann proposed that cold temperatures should favor the existence of large species, once their area-volume ratio allows them to conserve more heat than small species. As a consequence, a latitudinal cline of body sizes should emerge, a pattern latter named Bergmann's rule (Ashton *et al.*, 2000; Gaston *et al.*, 2008; Rodríguez *et al.*, 2008; Clauss *et al.*, 2013). Since the proposition of this rule, the pattern has been evaluated for a wide range of taxa, including endothermic and ectothermic organisms (Arnett & Gotelli, 1999; Drezner, 2003; Brehm & Fiedler, 2004; Olalla-Tárraga *et al.*, 2010; Clauss *et al.*, 2013).

In the macroecological literature, mammals are, like birds, an iconic example of Bergmann's rule (Blackburn *et al.*, 1999; Gaston *et al.*, 2008). Accordingly, body size latitudinal gradient has been confirmed by some authors studying intra-specific, cross-species and assemblage-based approaches (Ashton *et al.*, 2000; Meiri & Dayan, 2003; Diniz-Filho *et al.*, 2007; Clauss *et al.*, 2013; Santini *et al.*, 2017). However, meta-analytic

studies have not always concluded the existence of the rule (Ashton *et al.*, 2000; Freckleton *et al.*, 2003; Meiri & Dayan, 2003; Adams, 2008). In fact, these studies found a general application, but several taxa "broke the rule" (Ashton *et al.*, 2000; Freckleton *et al.*, 2003; Meiri & Dayan, 2003). Furthermore, the most recent and data-rich analysis at the intra-specific scale did not find evidence for a temperature-body size relationship (Riemer *et al.*, 2018). Conversely, recent comparative studies using cross-species approaches comprising almost all known mammal species concluded that mammals get larger at higher latitudes where cold temperatures dominate (Clauss *et al.*, 2013; Faurby & Araújo, 2016). Therefore, there is yet doubt about the existence of Bergmann's rule, even in the highly studied taxa such as mammals.

Commonly, studies use latitude, instead of temperature, as a surrogate variable to test Bergmann's rule. However, other environmental conditions such as productivity, climate stability, evapotranspiration and habitat availability, all have strong latitudinal gradients so they could explain body size variation (Geist, 1987; Blackburn *et al.*, 1999; Diniz-Filho *et al.*, 2007; Rodríguez *et al.*, 2008; Meiri, 2011). Indeed, studies evaluating body size variation within assemblage-based approaches found temperature might not be the only driver of body size cline (Diniz-Filho *et al.*, 2007, 2009; Rodríguez *et al.*, 2008). Nonetheless, until now, no comparative analyses using a cross-species approach have directly tested the relationship between body size and temperature in a mammalian class scale.

Testing Bergmann's rule, alongside all inferences about causal effects in macroecological and macroevolutionary studies, is a challenging endeavor as it makes use of observational data, whereas we cannot "reset the mammalian evolutionary tape" as thought by Stephen J. Gould (Gould, 1990). Therefore, Bergmann's rule researches are plagued by taphonomic process, research bias, species detectability, which results in

biased information toward some taxa and geographic regions (Nakagawa & Freckleton, 2008; Garamszegi & Moller, 2011; Gonzalez-Suarez *et al.*, 2012; Clavel *et al.*, 2014; Penone *et al.*, 2014). This biased information are translated into missing information, or knowledge shortfalls, about species ecology, trait, geographic occurrence and even about their own existence (Hortal *et al.*, 2015). Hence, the implication of biodiversity knowledge shortfall on evolutionary and ecological inferences are an ongoing research program (Diniz-Filho *et al.*, 2013; Hortal *et al.*, 2015; Oliveira *et al.*, 2016).

On the purpose to infer unbiased inferences from datasets with missing values, statistical literature has been publishing methods to deal missing data since 30's (Allan and Wishart 1930; Anderson 1957; Rubin 1976; see more in Little and Rubin 2002). Then, in 1976, Donald Rubin proposed a theory to analyze missing data based on assumptions about how data became missing through gathering procedure (Rubin, 1976). In his theory, if data were missing randomly, he named missing data mechanism as Missing Completely at Random (MCAR). On the other hand, if missing data is not random, but some variable within researcher database describes missing data probability, this mechanism is Missing at Random (MAR). Otherwise, when there is no variable explaining missing probability, the mechanism is Not Missing at Random (NMAR). Those missing data mechanism have practical implications, since researchers should choose methods proposed to deal with each mechanism, ensuring so, unbiased analyses (Rubin, 1976; Little & Rubin, 2002; Enders, 2010; van Buuren, 2012).

The use of appropriate methods regarding missing data has been advised in ecological literature (Fisher *et al.*, 2003; Nakagawa & Freckleton, 2008; Nakagawa, 2015), but few studies have implemented those methods into their research (Fisher *et al.*, 2003; Nakagawa & Freckleton, 2010; Fitzjohn *et al.*, 2014; Jetz & Freckleton, 2015; Bokma *et al.*, 2016; Uyeda *et al.*, 2017). Therefore, our goal here is to evaluate

Bergmann's rule hypothesis that temperature drives body size variation within the mammalian clade using a large and well-known database of mammalian traits, evaluating, in addition, the impact of ignoring missing data mechanisms in our inferences and parameter estimates.

MATERIAL AND METHODS

Database and phylogenies

To evaluate Bergmann's rule over a missing data perspective, we used PanTHERIA database (Jones *et al.*, 2009) as it is the most complete database about life history, ecology and geography of mammals, covering about 5400 extant and recently extinct mammal species. Moreover, PanTHERIA mirrors taxonomic and geographical biases of mammalian research, such as large, widely distributed mammals, inhabiting temperate regions are more prone to have data (Gonzalez-Suarez *et al.*, 2012). Thus, PanTHERIA is valuable to investigate how ignoring missing data influences macroecological researches.

To summarize environmental temperature along mammalian geographic range, we created a presence-absence matrix (PAM) of terrestrial mammals using International Union for Conservation of Nature (IUCN) polygons (International Union for Conservation of Nature, 2016) through *lets.presab* function of *letsR* package (Vilela & Villalobos, 2015). For each species, we extracted from WorldClim (Hijmans *et al.*, 2005) annual mean temperature (Bio1) in each inhabited grid cell and calculated the median temperature of those cells. Both Bio1 raster and PAM were previously transformed into equal area Mollweide projection with a resolution equivalent to 1º x 1º at the equator.

Evolutionary history was accounted by 101 phylogenies generated by Kuhn *et al.* (2011), which represents uncertainty on resolving polytomies throughout the mammalian species-level phylogeny published by Fritz *et al.* (2009). However, 397 species with geographic range or present in PanTHERIA were missing from phylogenies. We solved this by imputing missing species on phylogenies, allocating them at the most inclusive clade proposed by prior taxonomic and phylogenetic knowledge (Thomas *et al.*, 2013; Rangel *et al.*, 2015). Species were allocated within a monophyletic clade defined by their genus using function *add.species.to.genus* of *phytools* package (Revell, 2012). When species had no relatives on phylogenies, we searched on Open Tree of Life (Hinchliff *et al.*, 2015) for their most accepted locations. If one species was a sister lineage of a clade, we attached it at the stem branch of that clade. This imputation procedure used functions from *phytools* and *ape* (Paradis *et al.*, 2004; Revell, 2012) packages.

To allow compatibility of species among distribution range polygons, PanTHERIA and phylogenies, we checked and standardized synonymies to enhance species coverage (4309 species) in our subsequent analyses. We used *search_col* function in *taxize* package (Chamberlain & Szöcs, 2013) to search synonymies in PanTHERIA, phylogenies and IUCN polygons present on Catalog of Life database (Roskov *et al.*, 2017). Species missing on Catalog of Life had their synonymies accepted in accordance to the IUCN database.

Multiple imputation

To take missing data into account in our analyses, we applied Multiple Imputation by Chained Equation (MICE) (Buuren & Groothuis-Oudshoorn, 2011). We chose MICE to impute missing values, because of its simplicity and flexibility to deal with Missing at

Random (MAR) and Not Missing at Random (NMAR) mechanisms of missing values (van Buuren, 2012). MICE simulates iteratively posterior predictive distributions of missing values for each variable conditioned on observed data of other variables (van Buuren *et al.*, 2006; Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012). MICE usually assumes MAR mechanism while imputing values, but to access the sensibility of MAR assumption in our analyses, we also used Response Indicator model (Jolani, 2012), as it assumes NMAR mechanism and estimates missing data probability.

To impute missing values, we used life-history traits and phylogenetic information as regression predictors. To do so, we selected life-history traits with less than 80% of missing values (Table 1). Phylogenetic information was included in our model as phylogenetic eigenvectors (Diniz-Filho *et al.*, 1998). To obtain these eigenvectors, we decomposed a phylogenetic distance matrix of one of our phylogenies and extracted its eigenvectors, selecting for each life-history trait, the eigenvectors that eliminated their residual autocorrelation (Diniz-Filho *et al.*, 2012; Bauman *et al.*, 2018). All eigenvectors selected for all variables were included into imputation model. However, some variables were not well imputed by all eigenvectors and were excluded from analyses, since they did not have also much information to provide to the other variables (Table 1). Therefore, we chose litter size, neonate mass, body mass and temperature, besides those selected eigenvectors in our imputation model. Furthermore, we imputed missing values using Predictive Mean Matching (PMM), except for body mass. This algorithm predicts missing values based on a regression model, then selects, within observed values, the five closest values to that predicted one. Subsequently, the algorithm samples one value from

Table 1: Variables with less than 80% of missing values with their influx and outflux coefficients. Influx is the capacity of some variable to receive information from other

database variables. Outflux is the capacity of some variable to provide information to other database variables.

Variables	influx	outflux	missingness (%)
X5.1_AdultBodyMass_g	0.11	0.51	0.33
X13.1_AdultHeadBodyLen_mm	0.46	0.16	0.68
X9.1_GestationLen_d	0.46	0.07	0.73
X15.1_LitterSize	0.23	0.28	0.52
X5.3_NeonateBodyMass_g	0.55	0.05	0.78
X23.1_SexualMaturityAge_d	0.56	0.04	0.79
X25.1_WeaningAge_d	0.53	0.05	0.77
Temperature	0.02	0.94	0.04

Influx and outflux coefficients were proposed by (van Buuren, 2012).

that selected set of values (van Buuren, 2012). Therefore, PMM avoids unrealistic imputed values, but it is prone to eliminating phylogenetic signal. Thus, we imputed body mass by normal distribution, but constrained their values to minimum and maximum observed values.

We created 50 imputed datasets, for all imputation scenarios (MAR and NMAR), following van Buuren (2012) recommendation to generate as much imputations as average percentage of missing values in database variables. This is also in accordance with Graham *et al.* (2007) and von Hippel (2018) advices of imputation amount. MICE runs Monte Carlo Markov Chains (MCMC) to simulate imputations, then we run 30 iterates for MAR scenario and 50 iterates for NMAR (Response Indicator method), because Response Indicator method requires more iterations to estimate the probability of species misses values. We checked chains stabilization tracking mean and standard deviation traces of each variable (Fig.1, Supplementary Material). All imputations were run by *mice* package (Buuren & Groothuis-Oudshoorn, 2011).

Statistical analyses

We evaluated missing data effects on body mass and temperature relationship running separate analyses on the multiple imputed datasets described above. More specifically, we regressed a PGLS (Phylogenetic Generalized Linear Squares), with lambda transformation (Pagel, 1999; Freckleton *et al.*, 2002), between body mass and temperature in each of 50 created datasets using *phylolm* function in *phylolm* package (Tung Ho & Ané, 2014). Then, we pooled parameter estimates using Rubin's rule (Enders, 2010; van Buuren, 2012; Barnard & Rubin, 2018) to estimate unbiased parameters and 95 % confidence intervals.

To compare multiple imputation results to complete case analyses (i.e. excluding missing values), we ran PGLS, like abovementioned, over 101 phylogenies and pooled their results by Rubin's rule as it has been suggested by Nakagawa & de Villemereuil, (2015). Therefore, we could include both imputation and phylogenetic uncertainties into the same framework. All variables were log-transformed before both imputation and analyses, except temperature that was normalized by subtracting the mean values from its values and dividing it by its standard deviation. All analyses were run in R software 3.4 (R Core Team, 2017).

RESULTS

Our results did not support Bergmann's rule acting on mammalian clade, independently of the method used to deal with missing data, once the effect of temperature on body mass had 95% of probability to include absence of effect within their confidence intervals (Table 2). However, despite being not significant, we found opposite effects of temperature on body mass if analyses were run on complete case (negative effect) instead

of imputed datasets (positive effect). Furthermore, intercepts were lower for complete-case analyses in comparison with imputed analyses.

We also found body size residuals had high phylogenetic signal, but imputed datasets estimated lower phylogenetic signal ($\lambda = 0.89$) than complete case analyses ($\lambda = 0.97$). In addition, phylogenetic uncertainty (here a mixture of polytomies resolution and missing species imputation) influenced about 28% of temperature effect on complete-case analyses, but it had no effect on intercept estimates. Conversely, imputation uncertainty generated about 50% of temperature effect error and 20% of intercept error for both imputation scenarios (MAR and NMAR). It is worth noting that Response Indicator method (NMAR) results did not differentiate from MAR.

DISCUSSION

Bergmann's rule has been tested since it was first proposed about 170 years ago. Yet, it still raises questions about its existence in different major taxa, such as mammals. Here we did not find support for Bergmann's rule, an unexpected conclusion since it has been cited as hallmark of ecogeographic rules for a long time (Blackburn *et al.*, 1999; Gaston *et al.*, 2008). However, we did not obtain an uncommon result. Indeed, previous studies have already found evidences against Bergmann's rule while studying the whole mammalian clade (Clauss *et al.*, 2013; Faurby & Araújo, 2016). But, surprisingly, those studies concluded for the rule existence. For example, Clauss *et al.* (2013) concluded for appliance of the rule, but they found a latitude effect on body size of 0.002 g in log scale, in other words, the difference between a species at the latitude 90º and other species at latitude 0º would be 1.51 g, a difference comparable to measurement error. Thus, in fact, their results showed evidence against Bergmann's rule, not the opposite, as they

concluded. Faurby & Araújo (2016) also found that extant mammalian species do not follow the rule, but they argued that the absence of this pattern was a consequence of range contraction due to human activity and large mammal extinctions in temperate regions since the Pleistocene. Nevertheless, latitudinal effect in their study using historical ranges was comparable to Clauss *et al*. (2013) results. Therefore, studying the entire mammalian clade by cross-species approach seems to give no support for Bergmann' rule, probably a consequence of considering it as a homogeneous clade, with stationary body size evolution and constant effect of temperature on body size (Freckleton *et al.*, 2003; Meiri & Dayan, 2003)

Researches about tempo and mode of mammalian body size evolution have shown that mammalian evolution is not stationary, with changes in evolutionary rates through time, lineages and geography (Diniz-Filho *et al.*, 2007; Cooper & Purvis, 2010; Baker *et al.*, 2015; Clavel & Morlon, 2017). At the temporal scale, evolutionary rates increased in cold periods and decreased in warm times (Clavel & Morlon, 2017). Evolutionary rates also changed throughout the mammalian phylogeny (Cooper *et al.*, 2011; Venditti *et al.*, 2011; Baker *et al.*, 2015), increasing their rates in lineages becoming larger, so imprinting a directional evolution of body size toward larger values (Baker *et al.*, 2015), a pattern known as Depéret or Cope's rule (Alroy, 1998; Bokma *et al.*, 2016). Therefore, average evolutionary rates through time were a net outcome of lineages evolving at different speeds. In addition, similar patterns can be detected in geographical space, where evolutionary rates accelerated in cold regions and decreased in warmer regions (Cooper & Purvis, 2010). Taking all of these together, temperature seems to have influenced body size evolution, but as exemplified by Artiodactyla body size evolution (Carotenuto *et al.*, 2015), both space and time should be taken together to detect a clear effect of temperature, or latitude, on body size variation.

During the Quaternary period, global temperature decreased causing mammal migrations toward lower latitudes by tracking tropical climate reduction (Rolland *et al.*, 2018). Hence, most species conserved their "tropical niches" and accumulated at lower latitudes (Rolland *et al.*, 2018). It is plausible to suppose, given Bergmann's rule, that they also conserved their body sizes while migrating. Therefore, species inhabiting temperate regions might have evolved their niches to occupy colder temperatures, consequently increasing their body sizes. For instance, Freckleton *et al.* (2003) found that Bergmann's rule is detectable only for species larger than 160 g, then smaller species might have responded to climate changes by tracking their adequate temperature, or using other strategies such as increasing their fur insulation or burring behavior (Millien *et al.*, 2006; Porter & Kearney, 2009). Accordingly, temperature may indeed drive body size variation within the mammalian clade, but detecting that effect, we need to look at body size evolution as a complex and non-stationary process in which some species increase their sizes, while others decrease and most species conserve their ancestral body sizes in response to temperature changes. Therefore, we need to look for biological differences through a "biodiversity time" lens (Maddison & Fitzjohn, 2015), modelling their changes through time. If we consider the temperature effect on body size as a constant effect through all lineages, such as we did, the pattern becomes smoothed, so that the effect of temperature on body size becomes undetectable.

Although we have not found statistical support for temperature effect on body size, it is worth to note that different methods used to deal missing data generated inverse slope estimates. Complete-case analyses estimated a negative effect of temperature, which supports the Bergmann's rule statement. Conversely, imputation methods estimated a positive effect of temperature, which can be interpreted against the existence of Bergmann' rule. If species are not equally probable to have missing data, the MCAR

assumption underlying exclusion of missing data species before analyses is broken and, consequently parameter estimates become prone to be biased (Rubin, 1976; Nakagawa & Freckleton, 2008; Enders, 2010; van Buuren, 2012). Missing data pattern in PanTHERIA database has been mainly explained by species body size and distribution, thus missing data probability is not random, but inversely correlated to body size (Gonzalez-Suarez *et al.*, 2012). In fact, we found that species missing body sizes were at the left tail of the body size frequency, but they were, on average, larger than average sizes of species with data (Fig.2, Supplementary Material). Therefore, excluding missing species biased regression estimates toward lower intercepts and negative slopes than imputation analyses, potentially causing erroneous confirmation of Bergmann's rule.

As mentioned above, for imputation methods we estimated positive relationships between temperature and body size. This pattern might be an outcome of current largest species inhabiting only high temperature places (Fig.1) combined with imputed body size values that did not bias average body size estimates. Therefore, the detection of the small positive effect of temperature on body size at mammalian clade was possible due to imputation methods that guaranteed unbiased parameter estimates (Little & Rubin, 2002; Enders, 2010; van Buuren, 2012; Nakagawa, 2015; Murray, 2018). Hence, the recently increased interest about multiple imputation on ecological and evolutionary literature enhance the need of changing the practice of data exclusion for the use of proper methods to deal missing values (Nakagawa & Freckleton, 2010; Gonzalez-Suarez *et al.*, 2012; Penone *et al.*, 2014; Taugourdeau *et al.*, 2014).

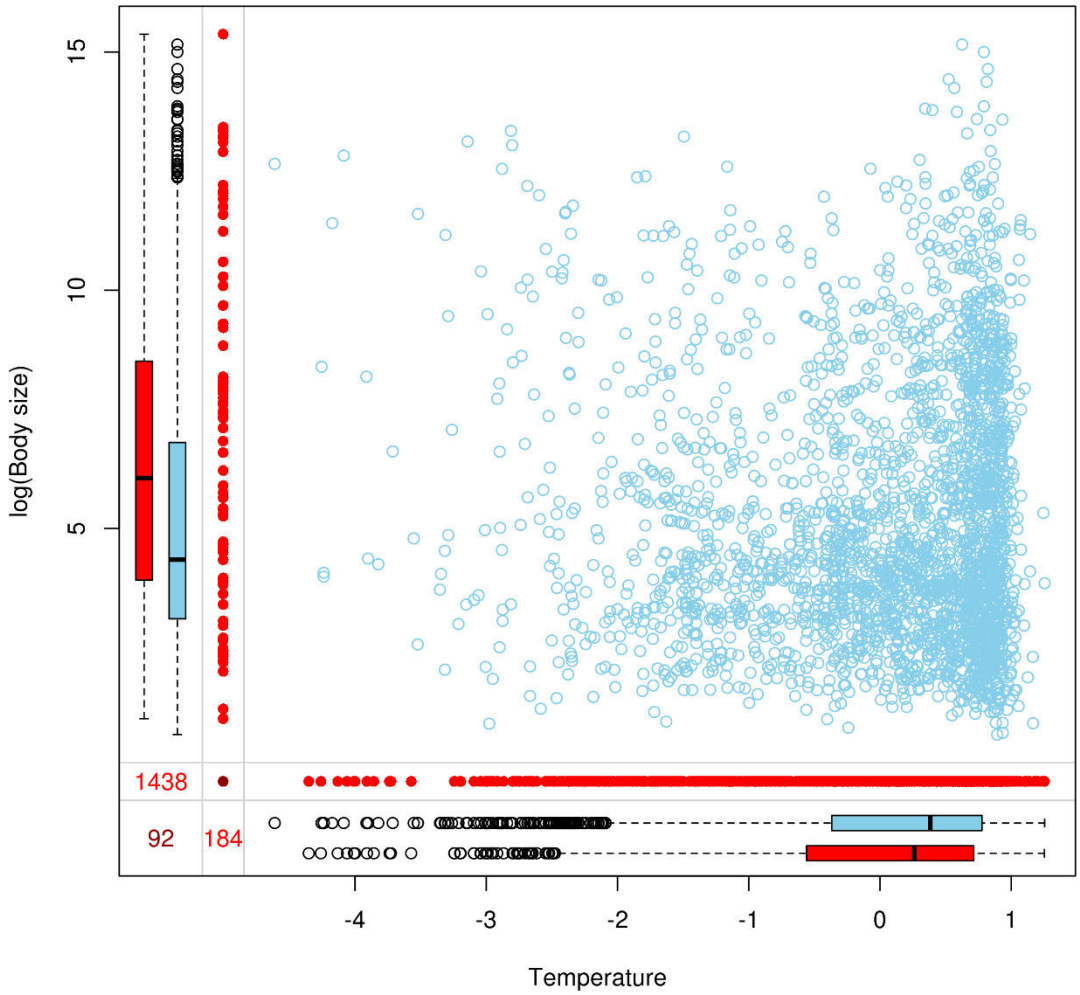


Figure 1: Scatter plot of temperature and body size. Blue boxes show variable distributions where both variables have observed values. Red boxes show variable distributions where the other variables has missing values. Red points indicates observed values of some variables in which other variable is missing values.

We applied two imputation methods, one assuming the MAR mechanism and the other the NMAR mechanism. Both methods estimated almost identical parameters. Thus, MAR assumption was plausible to describe how our data were missing (Jolani, 2012; van Buuren, 2012). These conclusions are not surprising, given that body size has high phylogenetic signal and well-established allometric relationship to several biological

features (Peters, 1983; Penone *et al.*, 2014). For instance, litter size and neonate mass were good variables to impute missing body sizes. Nonetheless, both traits had also missing values and then required imputations. Therefore, phylogenetic eigenvectors were indispensable as they were the only complete variables within the imputation process. Phylogenetic information has been for a long time suggested to predict missing values (Garland, Jr., & Ives, 2000; Bruggeman *et al.*, 2009; Guénard *et al.*, 2013; Swenson, 2014; Swenson *et al.*, 2016), because it carries out information about trait resemblance by shared ancestry as well as missing variables that makes species to evolve correlated.

However, to use phylogenetic eigenvectors, more studies should look for strategies to select eigenvectors, once through each imputation step, variable missing data are supposed to be correlated with different eigenvectors. We used a subset of eigenvectors that were selected for all variables, but it was not the optimal strategy once variables with a lot of missing data or with little phylogenetic signal generated values far from biological expectations, such as a litter size of 10^{-5}, unless we used PMM to impute. (Penone *et al.*, 2014) also found that eigenvectors could increase error sometimes using MICE, we suppose it was also due to eigenvector selection. Body mass, in turn, were modeled as Brownian motion, which has no constrain of possible values. Thus, body size could have imputed values out of known minimal and maximum mammalian sizes. We restricted imputation values to make the values realistic, but there is controversy about how to deal with outliers in multiple imputation literature (von Hippel, 2009; Rodwell *et al.*, 2014). Evolutionary models with bounds on trait evolution such as bounded Brownian Motion proposed by Boucher & Démery (2016) could improve phylogenetic imputation and this is probably a fruitful field to increase research.

Throughout our analyses, phylogenetic uncertainty showed little effect on parameter estimates, confirming that about 100 phylogenies, may be sufficient on

regression analyses that takes phylogenetic uncertainty into account (Nakagawa & de Villemereuil, 2015). Phylogenetic uncertainty was a mixture of polytomies resolution and missing species added on phylogenies. Therefore, species imputation seems to cause little impact on "phylogenetic regressions", which is in contrast to Rabosky (2015) results that proposed imputing species biased inferences. Probably, imputing species on phylogenies are dependent on the number of species being imputed (Rabosky, 2015), their phylogenetic depth and their phylogenetic relatedness, all of which can have large effects on analyses (Rangel *et al.*, 2015). Further studies could shed more light on how different methods impute species on phylogenies related to each other and their impacts on subsequent analyses. A possible strategy to add species on phylogenies could be to jointly estimate their phylogenetic positions, missing traits and model parameters (Slater *et al.*, 2012; Bokma *et al.*, 2016), then species would be located at phylogenetic positions that increase the fit of model parameters and probability of observing imputed data.

Multiple imputation had significant impact on our analyses, with almost 50 % of temperature effect error on body size was assigned to missing values uncertainty. To improve our inferences, we could decrease imputation variability by increasing the amount of imputations. However, we imputed the theoretical amount of imputations expected given our percentage of missing data (Graham *et al.*, 2007; van Buuren, 2012; von Hippel, 2018). Other alternative would be to include more variables into the imputation model, but the most complete mammalian database has most variables missing information for most species, despite the arduous effort expended by their developers (Gonzalez-Suarez *et al.*, 2012), so only litter size had some potential to provide information for imputation (Table 1). Our results, therefore, shows that ignoring missing values uncertainty, such as excluding missing values species or imputing values without

considering imputation uncertainty, may decrease parameter estimate errors, inflating type 1 errors.

Multiple imputation is a continuously developing statistical field (Murray, 2018), and has recently been commonly used in ecological and evolutionary studies (Fisher *et al.*, 2003; Nakagawa & Freckleton, 2010; Penone *et al.*, 2014; Taugourdeau *et al.*, 2014; Nakagawa & de Villemereuil, 2015). However, we have used methods developed in other scientific fields, such as medicine and psychology, which has different methodological demands in comparison to ecological and evolutionary studies. For example, Phylogenetic Comparatives methods (PCM) have been developed to deal with species non-independence and evolutionary model assumptions (Felsenstein, 1985; Hansen & Martins, 1996). Thus, incentives to develop PCM methods such as FitzJohn *et al.* (2009), Hadfield (2010) and Slater *et al.* (2012), which takes missing data into account could improve our research inference and design (Nakagawa, 2015), as values can be intentionally missing to balance data through phylogeny and geography, improving budget expenses and research bias (Nakagawa, 2015).

CONCLUSION

Macroecological literature has debated Bergmann' rule in mammals for a long time with controversial conclusions. Here we did not found support for that rule and suggest for testing the rule looking for shifts in body size evolution across time following temperature variation through lineages. Furthermore, missing values could bias parameter estimation and may affect inferences more than phylogenetic uncertainty. Thus, it is time for increasing research on phylogenetic imputation methods, both evaluating and proposing methods (Clavel *et al.*, 2014; Penone *et al.*, 2014; Paterno *et al.*, 2018) in order

to fulfill evolutionary and ecological demands, for example, methods regarding different evolutionary models, eigenvectors selection and phylogenetic clustering of missing values. Phylogenetic multiple imputation methods will enhance macroecological and macroevolutionary inferences as missing data constantly plague our researches.

ACKNOWLEDGMENTS

This work was developed in the context of National Institutes for Science and Technology (INCT) in Ecology, Evolution and Biodiversity Conservation, supported by MCTIC/CNpq (proc. 465610/2014-5). L.J. is supported by a CAPES Doctoral fellowship. JAFD.-F has been continuously supported by CNPq Productivity Grants.

REFERENCES

Adams, D.C. 2008. Phylogenetic meta-analysis. *Evolution.* **62**: 567–572.

Allan, F.E. & Wishart, J. 1930. A method of estimating the yeld of a missing plot in field experimental work. *J. Agric. Sci.* **20**: 399–406.

Alroy, J. 1998. Cope's rule and the dynamics of body mass evolution in north american fossil mammals. *Science.* **280**: 731–734.

Anderson, T.W. 1957. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Am. Stat. Assoc.* **52**: 200–203.

Araújo, M.B., Ferri-Yañes, F., Bozinovic, F., Marquet, P.A., Valladares, F. & Chown, S.L. 2013. Heat freezes niche evolution. *Ecol. Lett.* **16**: 1206–1219.

Arnett, A.E. & Gotelli, N.J. 1999. Geographic variation in life-history traits of the ant

lion, *Myrmeleon immaculatus*: evolutionary implications of Bergmann's rule. *Evolution.* **53**: 1180–1188.

Ashton, K.G., Tracy, M.C. & Queiroz, A. de. 2000. Is Bergmann's rule valid for mammals? *Am. Nat.* **156**: 390–415.

Baker, J., Meade, A., Pagel, M. & Venditti, C. 2015. Adaptive evolution toward larger size in mammals. *Proc. Natl. Acad. Sci.* **112**: 5093–5098.

Barnard, J. & Rubin, D.B. 2018. Sample degrees of freedom with multiple imputation. *Biometrika* **86**: 948–955.

Barnosky, A.D. 2001. Distinguishing the effects of the Red queen and Court Jester on Miocene mammal evolution in the northern Rocky Mountains. *J. Vertebr. Paleontol.* **21**: 172–185.

Bauman, D., Drouet, T., Dray, S. & Vleminckx, J. 2018. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. *Ecography.* **41**: 1–12.

Benton, M.J. 2009. The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time. *Science.* **323**: 728–732.

Blackburn, T.M., Gaston, K.J. & Loder, N. 1999. Geographic gradients in body size : a clarification of Bergmann 's rule. *Divers. Distrib.* **5**: 165–174.

Bokma, F., Godinot, M., Maridet, O., Ladevèze, S., Costeur, L., Solé, F., *et al.* 2016. Testing for Depéret's rule (body size increase) in mammals using combined extinct and extant data. *Syst. Biol.* **65**: 98–108.

Boucher, F.C. & Démery, V. 2016. Inferring bounded evolution in phenotypic characters from phylogenetic comparative data. *Syst. Biol.* **65**: 651–661.

Brehm, G. & Fiedler, K. 2004. Bergmann's rule does not apply to geometrid moths along an elevational gradient in an Andean montane rain forest. *Glob. Ecol. Biogeogr.* **13**: 7–14.

Bruggeman, J., Heringa, J. & Brandt, B.W. 2009. PhyloPars: Estimation of missing parameter values using phylogeny. *Nucleic Acids Res.* **37**: 179–184.

Buuren, S. van & Groothuis-Oudshoorn, K. 2011. mice : multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**.

Carotenuto, F., Diniz-Filho, J.A.F. & Raia, P. 2015. Space and time: The two dimensions of Artiodactyla body mass evolution. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **437**: 18–25.

Chamberlain, S.A. & Szöcs, E. 2013. taxize: taxonomic search and retrieval in R. *F1000Research* 1–26.

Clauss, M., Dittmann, M.T., Müller, D.W.H., Meloro, C. & Codron, D. 2013. Bergmann′s rule in mammals: A cross-species interspecific pattern. *Oikos* **122**: 1465–1472.

Clavel, J., Merceron, G. & Escarguel, G. 2014. Missing data estimation in morphometrics: How much is too much? *Syst. Biol.* **63**: 203–218.

Clavel, J. & Morlon, H. 2017. Accelerated body size evolution during cold climatic periods in the Cenozoic. *Proc. Natl. Acad. Sci.* **114**: 4183–4188.

Cooper, N., Freckleton, R.P. & Jetz, W. 2011. Phylogenetic conservatism of environmental niches in mammals. *Proc. R. Soc. B Biol. Sci.* **278**: 2384–2391.

Cooper, N. & Purvis, A. 2010. Body size evolution in mammals: Complexity in tempo and mode. *Am. Nat.* **175**: 727–738.

Diniz-Filho, J.A.F., Bini, L.M., Rangel, T.F., Morales-Castilla, I., Olalla-Tárraga, M.Á., Rodríguez, M.Á., *et al.* 2012. On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography.* **35**: 239–249.

Diniz-Filho, J.A.F., Bini, L.M., Rodríguez, M.Á., Rangel, T.F.L.V.B. & Hawkins, B.A. 2007. Seeing the forest for the trees: Partitioning ecological and phylogenetic components of Bergmann's rule in European Carnivora. *Ecography.* **30**: 598–608.

Diniz-Filho, J.A.F., Loyola, R.D., Raia, P., Mooers, A.O. & Bini, L.M. 2013. Darwinian shortfalls in biodiversity conservation. *Trends Ecol. Evol.* **28**: 689–95.

Diniz-Filho, J.A.F., Rodríguez, M.Á., Bini, L.M., Olalla-Tarraga, M.Á., Cardillo, M., Nabout, J.C., *et al.* 2009. Climate history, human impacts and global body size of Carnivora (Mammalia: Eutheria) at multiple evolutionary scales. *J. Biogeogr.* **36**: 2222–2236.

Diniz-Filho, J.A.F., Sant'Ana, C.E.R. & Bini, L.M. 1998. An eigenvector method for estimating phylogenetic inertia. *Evolution.* **52**: 1247–1262.

Drezner, T.D. 2003. Revisiting Bergmann's rule for saguaros (Carnegiea gigantea (Engelm.) Britt. and Rose): stem diameter patterns over space. *J. Biogeogr.* **30**: 353–359.

Enders, C.K. 2010. *Applied Missing Data Analysis*, 1st ed. New York, NY.

Faurby, S. & Araújo, M.B. 2016. Anthropogenic impacts weaken Bergmann's rule. *Ecography.* 1–2.

Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.* **125**: 1–15.

Fisher, D.O., Blomberg, S.P. & Owens, I.P.F. 2003. Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. *Proc. R. Soc. B Biol. Sci.* **270**:

1801–1808.

FitzJohn, R.G., Maddison, W.P. & Otto, S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst. Biol.* **59**: 458–464.

Fitzjohn, R.G., Pennell, M.W., Zanne, A.E., Stevens, P.F., Tank, D.C. & Cornwell, W.K. 2014. How much of the world is woody? *J. Ecol.* **102**: 1266–1272.

Freckleton, R.P., Harvey, P.H. & Pagel, M. 2002. Phylogenetic analysis and comparative data : a test and review of evidence. *Am. Nat.* **160**: 712–726.

Freckleton, R.P., Harvey, P.H., Pagel, M., Freckleton, R.P., Harvey, P.H. & Pagel, M. 2003. Bergmann's rule and body size in mammals. *Am. Nat.* **161**: 821–825.

Fritz, S. a, Bininda-Emonds, O.R.P. & Purvis, A. 2009. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecol. Lett.* **12**: 538–49.

Garamszegi, L.Z. & Moller, A.P. 2011. Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. *Syst. Biol.* **60**: 876–880.

Garland, Jr., T. & Ives, A.R. 2000. Using the past to predict the present: Confidence Intervals for regression equations in phylogenetic comparative methods. *Am. Nat.* **155**: 346–364.

Gaston, K.J., Chown, S.L. & Evans, K.L. 2008. Ecogeographical rules: elements of a synthesis. *J. Biogeogr.* **35**: 483–500.

Geist, V. 1987. Bergmann's rule is invalid. *Can. J. Zool.* **65**: 1035–1038.

Gonzalez-Suarez, M., Lucas, P.M. & Revilla, E. 2012. Biases in comparative analyses

of extinction risk: mind the gap. *J. Anim. Ecol.* **81**: 1211–22.

Gould, S.J. 1990. *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton, New York, NY.

Graham, J.W., Olchowski, A.E. & Gilreath, T.D. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* **8**: 206–213.

Guénard, G., Legendre, P. & Peres-Neto, P. 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. *Methods Ecol. Evol.* **4**: 1120–1131.

Hadfield, J.D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**: 1–22.

Hansen, T.F. & Martins, E.P. 1996. Translating between microevolutionary process and macroevolutionary patterns: correlation structure of interspecific data. *Evolution.* **50**: 1404–1417.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**: 1965–1978.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., *et al.* 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceeding Natl. Acad. Sci.* **112**: 12764–12769.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **46**: 523–549.

Huey, R.B., Kearney, M.R., Krockenberger, A., Holtum, J.A.M., Jess, M. & Williams,

S.E. 2012. Predicting organismal vulnerability to climate warming: roles of behaviour, physiology and adaptation. *Philos. Trans. R. Soc. B Biol. Sci.* **367**: 1665–1679.

International Union for Conservation of Nature. 2016. The IUCN Red List of Threatened Species. http://www.iucnredlist.org, Version 5. Accessed June 2016.

Jetz, W. & Freckleton, R.P. 2015. Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**: 20140016.

Jolani, S. 2012. Dual imputation strategies for analyzing incomplete data. University of Utrecht.

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S. a., O'Dell, J., Orme, C.D.L., *et al.* 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**: 2648–2648.

Kuhn, T.S., Mooers, A. & Thomas, G.H. 2011. A simple polytomy resolver for dated phylogenies. *Methods Ecol. Evol.* **2**: 427–436.

Little, R.J.A. & Rubin, D.B. 2002. *Statistical analysis with missing data*, 2nd editio. John Wiley & Sons, New Jersey.

Maddison, W.P.M. & Fitzjohn, R.G. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst. Biol.* **64**: 127–136.

Meiri, S. 2011. Bergmann's rule – what's in a name? *Glob. Ecol. Biogeogr.* **20**: 203–207.

Meiri, S. & Dayan, T. 2003. On the validity of Bergmann's rule. *J. Biogeogr.* **30**: 331–351.

Millien, V., Kathleen Lyons, S., Olson, L., Smith, F.A., Wilson, A.B. & Yom-Tov, Y. 2006. Ecotypic variation in the context of global climate change: Revisiting the rules. *Ecol. Lett.* **9**: 853–869.

Murray, J.S. 2018. Multiple imputation : A review of practical and theoretical findings. arXiv:1801.04058.

Nakagawa, S. 2015. Missing data: mechanisms, methods, and messages. In: *Ecological Statistics: Contemporary Theory and Application* (G. A. Fox et al., eds), pp. 81–105. Oxford University Press, Oxford, UK.

Nakagawa, S. & de Villemereuil, P. 2015. A simple and general method for accounting for phylogenetic uncertainty via Rubin's rules in comparative analysis. *PeerJ Prepr.* **3:e1216v1**: https://doi.org/10.7287/peerj.preprints.1216v1.

Nakagawa, S. & Freckleton, R.P. 2008. Missing inaction: the dangers of ignoring missing data. *Trends Ecol. Evol.* **23**: 592–596.

Nakagawa, S. & Freckleton, R.P. 2010. Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behav. Ecol. Sociobiol.* **65**: 103–116.

Olalla-Tárraga, M.Á., Bini, L.M., Diniz-Filho, J.A.F. & Rodríguez, M.Á. 2010. Cross-species and assemblage-based approaches to Bergmann's rule and the biogeography of body size in Plethodon salamanders of eastern North America. *Ecography.* **33**: 362-368.

Oliveira, U., Paglia, A.P., Brescovit, A.D., de Carvalho, C.J.B., Silva, D.P., Rezende, D.T., *et al.* 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Divers. Distrib.* 1–13.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature* **401**: 877–884.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

Paterno, G.B., Penone, C. & Werner, G.D.A. 2018. sensiPhy: an R-package for sensitivity analysis in phylogenetic comparative methods. *Methods Ecol. Evol.* **2018**: 1–7.

Penone, C., Davidson, A.D., Shoemaker, K.T., Marco, M. Di, Rondinini, C., Brooks, T.M., *et al.* 2014. Imputation of missing data in life-history traits datasets: which approach performs the best? *Methods Ecol. Evol.* **5**: 961–970.

Peters, R.H. 1983. *The ecological implications of body size*, 1st edition. Cambridge University Press, Cambridge.

Porter, W.P. & Kearney, M. 2009. Size, shape, and the thermal niche of endotherms. *Proceeding Natl. Acad. Sci.* **106**: 19666–19672.

Rabosky, D.L. 2015. No substitute for real data : A cautionary note on the use of phylogenies from birth – death polytomy resolvers for downstream comparative analyses. *Evolution.* **62**: 3207–3216.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput.* R Foundation for Statistical Computing, Vienna, Austria.

Rangel, T.F., Colwell, R.K., Graves, G.R., Fučíková, K., Rahbek, C. & Diniz-Filho, J.A.F. 2015. Phylogenetic uncertainty revisited: Implications for ecological analyses. *Evolution.* **69**: 1301–1312.

Revell, L.J. 2012. phytools: an R package for phylogenetic comparative biology (and

other things). *Methods Ecol. Evol.* **3**: 217–223.

Riemer, K., Guralnick, R.P. & White, E.P. 2018. No general relationship between mass and temperature in endothermic species. 1–16.

Rodríguez, M.Á., Olalla-Tárraga, M.Á. & Hawkins, B.A. 2008. Bergmann's rule and the geography of mammal body size in the Western Hemisphere. *Glob. Ecol. Biogeogr.* **17**: 274–283.

Rodwell, L., Lee, K.J., Romaniuk, H. & Carlin, J.B. 2014. Comparison of methods for imputing limited-range variables: A simulation study. *BMC Med. Res. Methodol.* **14**: 1–11.

Rolland, J., Silvestro, D., Schluter, D., Guisan, A., Broennimann, O. & Salamin, N. 2018. The impact of endothermy on the climatic niche evolution and the distribution of vertebrate diversity. *Nat. Ecol. Evol.*, doi: 10.1038/s41559-017-0451-9.

Roskov, Y., Abucay, L., Orrell, T., Nicolson, D., Bailly, N., Kirk, P.M., *et al.* 2017. Species 2000 & ITIS Catalogue of Life, 2016 Annual Checklist.

Rubin, D.. 1976. Inference and missing data. *Biometrika* **63**: 581–592.

Santini, L., González-Suárez, M., Rondinini, C. & Di Marco, M. 2017. Shifting baseline in macroecology? Unravelling the influence of human impact on mammalian body mass. *Divers. Distrib.* **23**: 640–649.

Slater, G.J., Harmon, L.J., Wegmann, D., Joyce, P., Revell, L.J. & Alfaro, M.E. 2012. Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate bayesian computation. *Evolution.* **66**: 752–762.

Swenson, N.G. 2014. Phylogenetic imputation of plant functional trait databases.

Ecography. **37**: 105–110.

Swenson, N.G., Weiser, M.D., Mao, L., Ara, M.B., Diniz-filho, A.F., Kollmann, J., *et al.* 2016. Phylogeny and the prediction of tree functional diversity across novel continental settings. *Glob. Ecol. Biogeogr.* **26:** 1–12.

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O. & Amiaud, B. 2014. Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data. *Ecol. Evol.* **4**: 944–958.

Thomas, G.H., Hartmann, K., Jetz, W., Joy, J.B., Mimoto, A. & Mooers, A.O. 2013. PASTIS: an R package to facilitate phylogenetic assembly with soft taxonomic inferences. *Methods Ecol. Evol.* **4**: 1011–1017.

Tung Ho, L.S. & Ané, C. 2014. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst. Biol.* **63**: 397–408.

Uyeda, J.C., Pennell, M.W., Miller, E.T., Maia, R. & McClain, C.R. 2017. The Evolution of energetic scaling across the vertebrate tree of life. *Am. Nat.* **190**: 185–199.

van Buuren, S. 2012. *Flexible Imputation of Missing Data*, 1st ed. Chapman and Hall/CRC, Boca Raton, Fl.

van Buuren, S., Brands, J.P.L., Groothuis-Oudshoorn, K. & Rubin, D.B. 2006. Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* **76**: 1049–1064.

Venditti, C., Meade, A. & Pagel, M. 2011. Multiple routes to mammalian diversity. *Nature* **479**: 393–396. Nature Publishing Group.

Vilela, B. & Villalobos, F. 2015. letsR: a new R package for data handling and analysis

in macroecology. *Methods Ecol. Evol.* **6**: 1229–1234.

von Hippel, P.T. 2018. How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociol. Methods Res.* 4912411774730.

von Hippel, P.T. 2009. How to impute squares, interactions, and other transformed variables. *Sociol. Methodol.* **39**: 265–291.

Supplementary Material

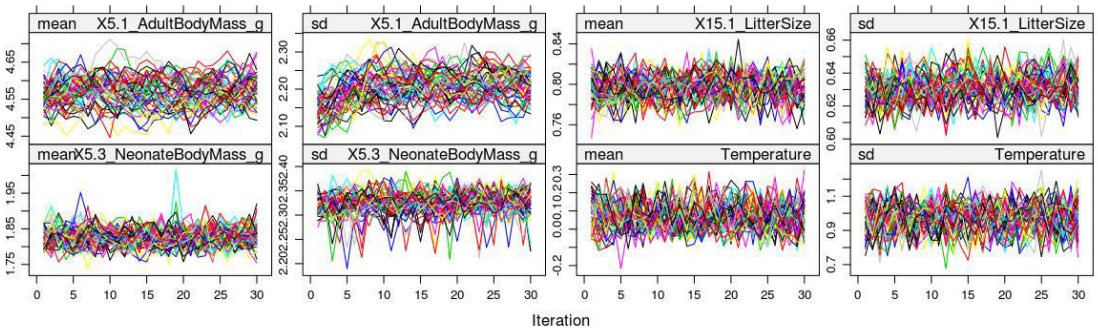


Figure 1: Trace line plots of mean and standard deviation (sd) for each independent Monte Carlo Markov Chain (MCMC). Line colors differentiate each 50 MCMC, convergence was achieved as statistical values (mean and sd) mixed through iteration and there was no tendency.

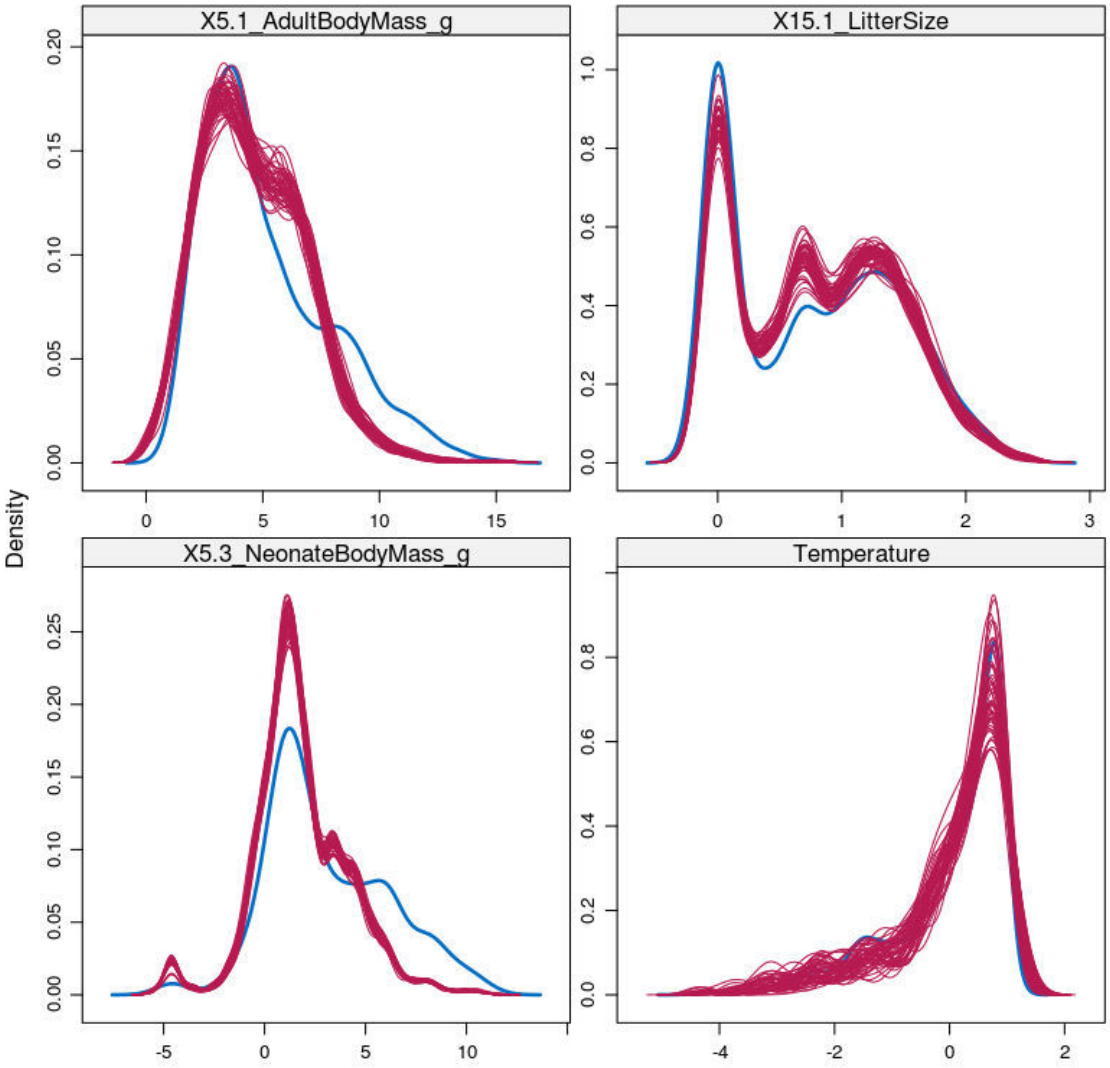


Figure 2: Histogram of imputed (magenta) and observed values (blue).

Conclusão geral

Nessa tese foi avaliada a imputação filogenética como técnica de preenchimento de dados, a sua aplicação e impactos em estudos macroecológicos. Nós apresentamos vantagens de sua aplicação, bem como suas limitações e ainda questões que necessitam de maior pesquisa.

No primeiro capítulo mostramos que a proporção e estrutura dos dados faltantes, conjuntamente com os modelos evolutivos dos atributos das espécies e os métodos utilizados para lidar com os dados faltantes, podem enviesar estudos macroecológicos. Por exemplo, ao estudarmos a estrutura filogenética dos atributos das espécies, o sinal filogenético pode ser estimado enviezadamente. Esse viés é, no entanto, dependente da metodologia utilizada para a sua estimação. Consequentemente, análises em bancos de dados imputados deveriam desconsiderar os valores imputados e modelar adequadamente os dados faltantes, por exemplo utilizando imputação múltipla, sempre em acordo com os objetivos e especificidades de cada estudo.

No segundo capítulo mostramos que primatas não seguem regra de ilha e ao predizermos o cérebro e massa do corpo de *Homo floresiensis*, constatamos que esses atributos não se desviam do que seria esperado pela história evolutiva de primatas. No entanto, nós não conseguimos descartar a possibilidade de ter havido efeito de ilha na evolução de *Homo floresiensis*, uma vez que nossos modelos superestimaram, em média, tanto a sua massa corpórea quanto o seu volume cerebral.

No terceiro capítulo encontramos que ao testar a regra de Bergmann em mamíferos, a desconsideração de dados faltantes pode enviesar a estimativa do efeito da temperatura na massa corpórea. No entanto, os dados faltantes não influenciaram a nossa conclusão de que a regra de Bergmann não se aplica aos mamíferos, quando as análises são realizadas para toda a classe.

Por fim, dados faltantes são uma regra em macroecologia. Portanto, a inclusão de métodos que consigam tratá-los adequadamente, no cotidiano dos macroecólogos, reduzirá os possíveis vieses sobre os processos que moldam os padrões de biodiversidade, assim como métodos espaciais e filogenéticos fizeram no passado. Essa tese, portanto, pode inspirar futuros estudos que busquem a melhor integração entre as teorias desenvolvidas para a lidar com dados faltantes, a imputação filogenética e a interpretação dos padrões macroecológicos.