Che-Castaldo, J., Jones, O.R., Kendall, B.E., Burns, J.H., Childs, D.Z., Ezard, T.H.G., Hernandez-Yanez, H., Hodgson, D.J., Jongejans, E., **Knight, T.**, Merow, C., Ramula, S., Stott, I., Vindenes, Y., Yokomizo, H., Salguero-Gómez, R. (2020):
Comments to “Persistent problems in the construction of matrix population models”
Ecol. Model. **416** , art. 108913

http://dx.doi.org/10.1016/j.ecolmodel.2019.108913

Title: Comments to “Persistent problems in the construction of matrix population models”

Published as:
Che-Castaldo, J., Jones, O. R., Kendall, B. E., Burns, J. H., Childs, D. Z., Ezard, T. H. G., Hernandez-Yanez, H., Hodgson, D. J., Jongejans, E., Knight, T., Merow, C., Ramula, S., Stott, I., Vindenes, Y., Yokomizo, H., & Salguero-Gómez, R. (2020). Comments to “Persistent problems in the construction of matrix population models”. Ecological Modelling, 416, [108913]. https://doi.org/10.1016/j.ecolmodel.2019.108913

Authors: Judy Che-Castaldo* (1), Owen R. Jones (2), Bruce E. Kendall (3), Jean H. Burns (4), Dylan Z. Childs (5), Thomas H.G. Ezard (6), Haydee Hernandez-Yanez (1), David J. Hodgson (7), Eelke Jongejans (8), Tiffany Knight (9), Cory Merow (10), Satu Ramula (11), Iain Stott (12), Yngvild Vindenes (13), Hiroyuki Yokomizo (14), & Roberto Salguero-Gómez (15)

(1) Alexander Center for Applied Population Biology, Conservation & Science Department, Lincoln Park Zoo, Chicago, IL 60614-4712 USA
(2) Interdisciplinary Center on Population Dynamics & Department of Biology, Campusvej 55, University of Southern Denmark, 5230 Odense M, Denmark
(3) Bren School of Environmental Science & Management, University of California Santa Barbara, Santa Barbara, CA 93106-5131 USA
(4) Department of Biology, Case Western Reserve University, Cleveland, OH 44106-7080 USA
(5) Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, UK
(6) Ocean & Earth Science, National Oceanography Centre Southampton, University of Southampton Waterfront Campus, Southampton SO14 3ZH, UK
(7) Centre for Ecology & Conservation, College of Life and Environmental Sciences, University of Exeter, Cornwall, UK
(8) Animal Ecology and Physiology Department, Radboud University, Nijmegen, The Netherlands
(9) Institute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, 06108 Halle (Saale), Germany; Department of Community Ecology, Helmholtz Centre for Environmental Research-UFZ, Theodor-Lieser-Straße 4, 06120 Halle (Saale), Germany; German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany
(10) Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA
(11) Department of Biology, University of Turku, FI-20014, Turku, Finland
(12) School of Life Sciences, University of Lincoln, Brayford Pool, LN6 7TS, Lincoln UK
(13) Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway
(14) Center for Health and Environmental Risk Research, National Institute for Environmental Studies, Tsukuba, 305-8506, Japan

(15) Department of Zoology, 11a Mansfield Rd, University of Oxford, OX1 3SZ, Oxford, UK

*Corresponding author: jchecastaldo@lpzoo.org; 301-351-8290

The use of matrix population models (MPMs) to summarize and analyze the empirically observed demography of both plant and animal populations has exhibited rapid growth in recent decades (Salguero-Gómez et al. 2015, Salguero-Gómez et al. 2016). Comparative analyses using independent collections of MPMs have generated valuable ecological insights (e.g., Franco & Silvertown 1990, Franco & Silvertown 2004, Ramula et al. 2008, Buckley et al. 2010, Burns et al. 2010, Salguero-Gómez & Casper 2010, Stott et al. 2010, Crone et al. 2011, Zeigler et al. 2013), leading to the recognition that a comprehensive MPM dataset would be a valuable resource for the ecological research community. We therefore launched the COMPADRE Plant Matrix Database (Salguero-Gómez et al. 2015) and the COMADRE Animal Matrix Database (Salguero-Gómez et al. 2016; http://compadre-db.com/), which comprise 1204 plant and animal species and 10949 MPMs as of October 2019.

Our initial goal with COMPADRE and COMADRE was to present MPMs as intended by the authors of the original publications, relying on the authors and the peer review process to ensure that the published population projection matrix correctly described the intended life history. Thus, our error-checking procedure to date has focused for the most part on eliminating typographical errors introduced during digitization, rather than errors of MPM construction made by the original authors. However, Kendall et al. (2019) have shown that a substantial fraction of animal MPMs in peer-reviewed papers contain errors in construction, such that model components are missing and/or the model does not accurately represent the described life history of the species. Plant MPMs are also prone to errors in construction, including dormant life stages (most commonly a seed bank) that are missing (Doak et al. 2002, Nguyen et al. 2019) or incorrectly incorporated, leading to a one-year delay in the life cycle (Caswell 2001, p. 60). Although some construction errors may not dramatically affect estimates of demographic parameters in some cases, they can be crucial in others. Estimates of generation time and population growth rate, in particular, can be sensitive to missing components or life stages (Kendall et al. 2019, Nguyen et al. 2019). Construction errors may also lead to directional biases, such as underestimating the elasticities of population growth rate to survival when fertility coefficients do not include survival (Kendall et al. 2019), and these biases may affect results in comparative analyses. Moving forward, there is a clear need for better training materials to help ecologists and conservation scientists (many with little experience in mathematical modeling) construct MPMs from their data. But how can we ensure that previously published MPMs, including those already compiled in COMPADRE and COMADRE and those remaining to be digitized, correctly represent the intended life histories, and thus support meaningful synthetic analyses?

We believe that an important principle in big data analyses can be summed up as *"Just because you can, does not mean you should"*. That is, although large volumes of data and tools that enable their fast analysis are available, they should not be used without careful consideration and evaluation. There are many possible sources of error whenever large datasets are compiled, and we have previously discussed some of these (including errors in matrix construction by authors) as a cautionary note to data users (Salguero-Gómez et al. 2016). We also maintain an open, frequently updated digitization protocol, and an extensive user guide (both available on our GitHub repository, https://github.com/jonesor/compadreDB). These open, transparent policies are essential for users to fully understand the data and choose the subsets of MPMs that are most appropriate for addressing their specific research questions.

In our collective cases, comparative analyses of demographic data from COMPADRE and COMADRE (e.g., Vindenes and Engen 2017, Beckman et al. 2018, Che-Castaldo et al. 2018, Davidson et al. 2019, Healy et al. 2019, Jones et al. 2019, Nguyen et al. 2019, Roper et al. 2019, Pistón et al. 2019) have each resulted in a separate sample size because the research questions required a different set of selection criteria. For instance, for some questions related to estimating various moments of longevity, the subcomponents of fertility are not necessary (if one assumes no trade-offs between reproduction and survival), and so MPMs that do not include those (indicated by the metadata field "MatrixFec") can be used. Similarly, to calculate the population growth rate from an MPM, the subcomponents of survival, sexual reproduction and clonality do not need to be separated (indicated by the metadata field "MatrixSplit"), but to estimate life history traits and vital rate sensitivities/elasticities, it is fundamental that they are separated. Thus, the set of selection criteria used for a specific study must be carefully designed before the analysis begins. Moreover, the availability of digitized data should not be a substitute for examining the original sources, which may provide important details of interest for the user's specific goals. For this reason (and to facilitate the citation of original sources), COMPADRE and COMADRE include bibliographic information for the source studies.

Nevertheless, we agree that addressing the newly-identified MPM construction errors in COMPADRE and COMADRE would be a useful service for the scientific community. In fact, in recent years we have started to include information on data quality in the metadata. For example, there is a field ("SurvivalIssue") for noting when the summed survival for a given age/stage exceeds 1.0, which may indicate a typographical error or a sensitivity matrix being incorrectly identified as an MPM in the original source. Other issues, such as when GPS coordinates are approximate rather than accurate, or environmental conditions that, while naturally occurring, may not be representative of typical conditions for a given population (e.g., fires, herbivory, droughts), are noted in the unstructured "Observations" field. In the coming months, we will take further steps including updating the metadata to record (1) whether the MPM is based on a pre-breeding or post-breeding census whenever the information is available, and (2) whether the MPM in the database has been altered from the original published one and why [e.g., typo in the original source; inclusion of an unnecessary seed stage where no permanent seed bank exists (Caswell 2001, p. 60)], or whether the MPM was calculated from other published data along with reference to those data (e.g., a life table, integral projection model, individual-based model). We have updated the digitization protocol to include Kendall et al. (2019) as essential reading so that our team can avoid the identified common errors when constructing MPMs from published life cycle data. Future releases of the database will include a unique identification number for each existing and new MPM, so that users can look up a specific MPM and all of its associated observations and metadata. They can also share with the community any additional metadata they collect for their own analyses, which could be linked back to the full dataset via the unique ID.

In the longer term, our plans include identifying MPMs that are likely to have construction issues following the protocol presented in Appendix B of Kendall et al. (2019). If the publication contains sufficient data, we will attempt to correct the MPMs and indicate why and how we have done so. This way, users will be able to identify the modified matrices and determine whether those modifications are appropriate for their use. In the last year, we have transitioned to a relational database that will allow us to keep a record of changes to the MPM data, so that we can record the MPMs as presented by original authors and also update them to fix known errors such as those mentioned above. However, often the published paper and supplementary materials do not include the necessary information (e.g., a separate life table or mean development time for each life history stage) to help us contextualize the MPMs and determine specific construction methods. Our digitization team invests a significant amount of time contacting authors for clarification and additional information. Without the original researchers' input, there may be many MPMs that we cannot correct or even flag with potential issues. This highlights the need for authors to consider providing as much detail on the life cycles, data collection, and biogeographic context for their species and populations as possible to enable reproducible research.

Although finding errors like the ones reported by Kendall et al. (2019) is worrisome, we emphasize that an evaluation of data quality issues and their extent (not to mention large-scale assessments in general) would not be possible without the existence of a database for MPMs. A database also allows us to implement solutions systematically across the entire set of demographic studies represented in the database. As a centralized repository of MPM data, COMPADRE and COMADRE can: (1) serve as a point of reference for recording whether a particular publication may have miscalculated MPMs, (2) pool efforts from independent groups or individuals to identify and/or fix the miscalculated MPMs, and (3) enable users to access and benefit from previously corrected MPMs. The database will only improve as we continue to aggregate data, assess data quality, and make corrections. Moreover, the mission of the COMPADRE and COMADRE team goes beyond releasing large volumes of data, and includes teaching workshops around the globe about MPMs and engaging the demographic community. We can therefore use our existing platform, including educational materials on our GitHub repository (see link above) and our regular user engagement workshops, to help raise awareness and promote more careful construction of MPMs.

We encourage and welcome the community to help validate the MPM data in COMPADRE and COMADRE, as the databases are currently supported by limited funding and volunteered researcher time. Conducting a comprehensive check in the databases for construction errors will require substantial effort, and we are only beginning to understand the extent of these issues. We are developing website tools (such as the new function for reporting data errors and corrections) to gather input and facilitate collaborative data validation. Participation from users and original authors will be vital in this effort to help us deliver reliable, open-access MPM data for demographic researchers and the broader scientific community.

Literature Cited

Buckley, Y. M. et al. 2010. Causes and consequences of variation in plant population growth rate: a synthesis of matrix population models in a phylogenetic context. *Ecology Letters* 13, 1182–1197.

Burns, J. H. et al. 2010. Empirical tests of life-history evolution theory using phylogenetic analysis of plant demography. *Journal of Ecology* 98, 334–344.

Beckman N, Bullock J, Salguero-Gómez R. 2018. High dispersal ability is related to fast life history strategies. *Journal of Ecology* 4, 1349–1362.

Caswell, H. 2001. *Matrix Population Models: Construction, Analysis, and Interpretation*, 2nd edn. Sinauer Associates, Inc, Sunderland, MA, USA

Che-Castaldo, J., Che-Castaldo, C., Neel, M. C. 2018. Predictability of demographic rates based on phylogeny and biological similarity. *Conservation Biology* 32, 1290–1300.

Crone, E. E. et al. 2011. How do plant ecologists use matrix population models? *Ecology Letters* 14, 1–8.

Davison R., Stadman M., Jongejans E. 2019. Stochastic effects contribute to population fitness differences. *Ecological Modelling* 408:108760.

Doak, D.F., Thomson, D., Jules, E.S. 2002. *Population viability analysis for plants: Understanding the demographic consequences of seed banks for population health.* in: Beissinger S.R., McCullough D.R. (eds.) Population viability analysis. The University of Chicago Press, Chicago, Illinois, USA.

Franco, M. & Silvertown, J. 1990. Plant demography: What do we know? *Evolutionary Trends in Plants* 4, 74–76.

Franco, M. & Silvertown, J. 2004. A comparative demography of plants based upon elasticities of vital rates. *Ecology* 85, 531–538.

Healy K., Ezard T., Jones O., Salguero-Gómez R., Buckley Y. 2019. Animal life history is shaped by the pace of life and the distribution of age-specific mortality and reproduction. *Nature Ecology & Evolution* 3, 1217–1224.

Jones O.R., Ezard T.H.G., Dooley C., Healy K., Hodgson D.J., Mueller M., Townley S., Salguero-Gomez R. 2019. My family and other animals: Human demography under a comparative cross-species lens. In "Human Evolutionary Demography", Burger O., Lee R., & Sear R. (eds). Open Book Publishers. Cambridge. UK.

Kendall, B.E., Fujiwara, M., Diaz-Lopez, J., Schneider, S., Voigt, J., Wiesner, S. 2019. Persistent problems in the construction of matrix population models. *Ecological Modelling* 406, 33–43 https://doi.org/10.1016/j.ecolmodel.2019.03.011

Nguyen V., Buckley Y.M., Salguero-Gómez R., Wardle G.M. 2019. Consequences of neglecting cryptic life stages from demographic models. *Ecological Modelling* 408, 108723.

Pistón N., de Bello F., Dias A., Götzenberger L., de Mattos E., Rosado B., Salguero-Gómez R., Carmona C. 2019. Multidimensional ecological analyses demonstrate how interactions between functional traits shape fitness and life history strategies. *Journal of Ecology* 107: 2317–2328.

Ramula, S., Knight, T. M., Burns, J. H. & Buckley, Y. M. 2008. General guidelines for invasive plant management based on comparative demography of invasive and native plant populations. *Journal of Applied Ecology* 45, 1124–1133.

Roper M., Capdevila P., Salguero-Gómez R. 2019. Senescence - still an unsolved problem of biology. Biorxiv: DOI 10.1101/739730

Salguero-Gómez, R. & Casper, B. B. 2010. Keeping plant shrinkage in the demographic loop. Journal of Ecology 98, 312–323.

Salguero-Gómez, R. et al. 2015. The COMPADRE Plant Matrix Database: an open online repository for plant demography. *Journal of Ecology* 103:202–218

Salguero-Gómez, R. et al. 2016. COMADRE: a global data base of animal demography. *Journal of Animal Ecology* 85:371–384

Stott, I., Franco, M., Carslake, D., Townley, S. Hodgson, D. 2010. Boom or bust? A comparative analysis of transient population dynamics in plants. *Journal of Ecology* 98, 302–311.

Vindenes, Y., Engen, S. 2017. Demographic stochasticity and temporal autocorrelation in the dynamics of structured populations. *Oikos* 126, 462–471.

Zeigler, S., Che-Castaldo, J. Neel, M. 2013. Actual and potential use of population viability analysis in recovery of plant species listed under the U.S. Endangered Species Act. *Conservation Biology* 27, 1265–1278.