Response to Valle and Zorello Laporta: Clarifying the use of instrumental variable methods to understand the effects of environmental change on infectious disease transmission

Running head: Response to the critique by Valle & Zorello Laporta

Andrew J. MacDonald[1*] & Erin A. Mordecai[2]

1. Earth Research Institute and Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, USA

2. Department of Biology, Stanford University, Stanford, CA, USA

*Corresponding author: Bren School of Environmental Science and Management, University of California, Santa Barbara, CA 93106-5131; andy.j.macdon@gmail.com

Keywords: causal inference; instrumental variables; environmental change; vector-borne disease; earth observation data

Abstract: 164 words

Text: 2264 words

Figures: 0

Tables: 0

SI: 1 SI table

Abstract

Identifying the effects of environmental change on the transmission of vector-borne and zoonotic diseases is of fundamental importance in the face of rapid global change. Causal inference approaches, including instrumental variable (IV) estimation, hold promise in disentangling plausibly causal relationships from observational data in these complex systems. Valle and Zorello Laporta recently critiqued the application of such approaches in our recent study of the effects of deforestation on malaria transmission in the Brazilian Amazon on the grounds that key statistical assumptions were not met. Here, we respond to this critique by: 1) deriving the IV estimator in order to clarify the assumptions that Valle and Zorello Laporta conflate and misrepresent in their critique; 2) discussing these key assumptions as they relate to our original study and how our original approach reasonably satisfies the assumptions; and 3) presenting model results using alternative instrumental variables that can be argued more strongly satisfy key assumptions, illustrating that our results and original conclusion—that deforestation drives malaria transmission—remain unchanged.

Main Text

There is substantial and increasing interest in understanding the role that processes of global change are playing in the ecology and transmission of vector-borne and zoonotic diseases.[1,2] While these questions are of fundamental importance given the increasing rate of climate and land use change, and the large proportion of emerging infectious diseases that are vector-borne or of zoonotic origin,[3] causally linking these two processes is an enormous challenge. Take as an example the case of deforestation impacts on malaria transmission in the Brazilian Amazon, the focus of MacDonald & Mordecai[4] and the critique by Valle & Zorello Laporta.[5] The gold standard of a randomized controlled trial in which deforestation is experimentally manipulated and randomly assigned to different regions to assess its impact on malaria transmission presents obvious logistical and ethical barriers that make such an approach largely infeasible. As a result, researchers must rely on observational data and employ statistical approaches to approximate, as closely as possible, the experimental ideal.

One promising set of statistical techniques—broadly referred to as causal inference methods, which includes Instrumental Variable (IV) estimation, are increasingly being leveraged to disentangle plausibly causal relationships from observational data in ecology. Due to the challenges described above, these approaches have been employed by researchers assessing global change impacts on infectious disease,[6-14] including in another recent study investigating the effects of deforestation on malaria transmission in Brazil,[14] with similar results to our own work. Valle and Zorello Laporta[5] rightly point out that model assumptions are critically important in such approaches, and that causal conclusions should be carefully drawn in these contexts. However, the authors unfortunately conflate the assumptions of IV estimation in their perspective piece. As a relatively new approach in ecology and environmental science,[6] it is important that the underlying assumptions are clear for appropriate application.

IV is a useful approach to overcome what is known as endogeneity bias, which is due to a relationship between the error term and one or more of the explanatory variables, (formally, $E[\varepsilon_i|x_i] \neq 0$ where ε and x represent the error term and explanatory variable for observation i). Such a relationship could be due to bidirectional causality where, for example, deforestation may drive malaria transmission but malaria burden may also influence rates of deforestation. In IV, a third variable, known as an instrument (z_i), is used to isolate exogenous variation in explanatory variable x_i and recover a statistically consistent estimator for the true relationship between the exogenous variable and the outcome.

The instrument must meet two conditions for IV to be a consistent estimator, which are sometimes termed "relevance" and "exclusion" criteria. In words, the instrument must be statistically associated with the endogenous variable ("relevance") and must be related to the outcome only through its relationship with the endogenous variable ("exclusion"). While the wording is easy to remember, it leaves much open to interpretation. For example, does relevance require a causal link? Does exclusion require statistical independence? The derivation makes these key assumptions much more apparent. Before showing the derivation, we will first provide brief background to our original study,[4] the critique by Valle & Zorello Laporta[5] and our response.

In MacDonald & Mordecai,[4] we were first interested in predicting annual malaria incidence as a function of annual deforestation, and use aerosol optical depth (AOD) in the month of September from MODIS satellite imagery as our "instrument." We expand on the methodology and terminology below, but set the context of the argument here. Valle & Zorello

Laporta[5] have two critiques of our IV approach. The first, however, is a misrepresentation of the assumptions of IV, namely that a valid IV requires that the IV has a *causal* effect on the endogenous explanatory variable. They state, “However, it is deforestation that causes aerosol pollution […] rather than aerosol pollution that causes deforestation […] As a result, [the relevance] assumption is clearly violated.” As we show below, causality is not required.[15] Rather, there must be an “association”, or more specifically, the covariance between the instrument and the endogenous variable must not be zero. However, it is possible that an instrumental variable itself introduces endogeneity bias if it does not meet the exclusion criteria, and this can be particularly problematic in the case of “weak instruments” as we show below. This can occur, for example, in cases where the instrument (e.g., AOD) is strongly driven by the endogenous predictor variable (e.g., deforestation). In our case, we chose AOD as an instrument for deforestation, as it is an indicator of human activity on the landscape.[16] Further, over our study period, AOD was decoupled from deforestation as biomass burning in the Brazilian Amazon—and resulting AOD—was primarily driven by fires intentionally set to keep *existing* pastures and agricultural lands clear[16] and by drought conditions leading to wildfires in already degraded forests,[16-18] rather than by new deforestation activity.

Nevertheless, to explore the extent to which our original IV estimates of the effect of deforestation on malaria may have been affected by potential endogeneity introduced by the use of AOD as an IV, we run additional IV models using 1) last year’s AOD as an instrument for this year’s deforestation, and 2) remotely sensed, average municipality soil quality[19] processed in Google Earth Engine,[20] interacted with annual international soy and beef commodity prices from the World Bank. We chose last year’s AOD since it is correlated with this year’s deforestation (relevance), but this year’s deforestation could not have caused last year’s AOD. While this addresses the issue of reverse causality, it is plausible that there remain endogeneity issues in this context. For example, if last year's AOD somehow acts upon this year's malaria through mechanisms beyond deforestation, then the exclusion criteria would fail. To address these potential lingering concerns, we run additional models using soil quality coupled with international agricultural commodity prices for key Brazilian exports, which may influence a land owners' decision to clear forest for agricultural production (relevance); in this case, deforestation rates do not cause soil quality and are highly unlikely to shift international commodity prices (exclusion). We run these IV models on our interior Amazon sample of municipalities, where active deforestation rates are highest and where we predict forest clearing should have the strongest effect on malaria transmission,[4] predicting both total malaria and *Plasmodium falciparum* malaria incidence, following our original study.[4] Results are presented in the SI (Table S1). In brief, we find significant positive effects of deforestation on malaria transmission in each of these additional model specifications, with coefficients of similar, though slightly larger magnitude than our original study. Our main conclusion, that deforestation increases malaria transmission in the Brazilian Amazon, remains unchanged.

The second goal of MacDonald & Mordecai[4] is to understand whether annual malaria burden feeds back to influence annual rates of deforestation, and we use optimal temperature for malaria transmission in the dry season as our instrument for malaria. Optimal temperature was defined as the sum of days falling within a narrow temperature band that is optimal for malaria transmission (24-26ºC) based on earlier mosquito and parasite trait-based mechanistic modeling studies.[21] Valle & Zorello Laporta's[5] second critique is that the exclusion assumption may be violated in this model because "it is possible that temperature affects deforestation not only through malaria, but also through other causal paths," particularly the relationship between temperature and agricultural gross domestic production.[22] In other words, favorable temperatures for mosquitos and malaria parasites may affect deforestation not just through malaria, but by also being favorable agricultural growing conditions, which increase the potential value of forest clearing. We agree that temperature is important to both agriculture and malaria, and that those clearing land may consider the land's growing potential. However, rather than counting the number of days in a 2ºC temperature window during the dry season, we suggest agricultural producers will instead consider the general growing conditions of a region as it relates to commonly grown crops—for example, soil quality, climate, topography, and infrastructure. As land clearing for agriculture is a large and long-term investment, average growing conditions are much more likely to influence clearing decisions than are small deviations in weather from year to year.

There are two additional primary reasons that our IV, optimal malaria transmission temperature, is highly unlikely to fail the exclusion criteria. First, we specifically employ municipality "fixed effects" or dummy variables[15] to remove roughly time invariant characteristics specific to each municipality through differencing. Thus, average characteristics (e.g., soil quality, average precipitation, average temperature) that are likely to influence the evolution of regional agricultural land use and the location of processing plants and other infrastructure are removed and the model is identified from deviations from the municipality-specific mean. Second, the range of optimal average temperatures for soybean—Brazil's main crop by area and production[23]—cultivation and development in Brazil is from 20ºC to 35ºC.[24] Recall optimal temperature for malaria transmission is 24ºC to 26ºC, and we use the number of days in the dry season within this narrow temperature band as our instrument. Thus, an additional day at 25ºC relative to 27ºC would be expected to lead to increases in malaria transmission. However, this same change in temperature would likely have a trivial impact on soy yields, as both temperatures are well within the bounds of optimal soy cultivation. Given the breadth of favorable temperatures for soy, it is unlikely that changes in the number of days between 24ºC to 26ºC will influence land clearing decisions for agricultural production.

We too feel that causal inference approaches hold much promise in disease ecology, and agree that researchers interested in exploring the use of such methods should carefully consider model assumptions. Toward that end, we briefly derive the simplest form of IV to illustrate to potential users what is under the hood of the IV approach and how the exclusion and relevance assumptions function in this technique.

Deriving the IV Estimator: To keep it as intuitive as possible, let us assume a bivariate regression of the form,

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad 1$$

Where y_i is the outcome variable (e.g., malaria incidence) for observation (e.g., municipality) i, x_i is the endogenous explanatory variable (e.g., deforestation), ε_i is the error term, α is the intercept, and β is the coefficient of interest.

To derive the IV estimator, we can take the covariance of each side of equation 1 with respect to the instrument, z_i:

$$cov(z_i, y_i) = cov(z_i, \alpha) + cov(z_i, \beta x_i) + cov(z_i, \varepsilon_i) \qquad 2$$

$$= 0 + \beta cov(z_i, x_i) + \; cov(z_i, \varepsilon_i) \qquad 3$$

Since α is a constant, and the covariance of a variable with a constant is 0, the first term drops out. Similarly, because β is a constant, it can be removed from the covariance. The exclusion assumption of IV is that the instrument (z_i) only affects the outcome through changes in the endogenous variable (x_i), which is more formally written as $cov(z_i, \varepsilon_i) = 0$. Thus with basic rearranging, we have derived the IV estimator (β_{IV}),

$$\beta_{IV} = \frac{cov(z_i, y_i)}{cov(z_i, x_i)} \; . \qquad 4$$

Consistency of IV: If we then want to illustrate that the IV estimator is consistent—in other words, as the sample size gets larger and larger the distribution of the estimator converges to the true parameter value—we can plug the right-hand side of equation 1 into y_i in equation 4. We substitute β_{IV} with $\widehat{\beta_{IV}}$ since we are considering whether the estimated slope from an IV converges in probability to the true slope β.

$$plim \; \widehat{\beta_{IV}} = \frac{cov(z_i, \alpha + \beta x_i + \; \varepsilon_i)}{cov(z_i, x_i)} \; . \qquad 5$$

Following a similar logic as with equation 3, equation 5 becomes:

$$plim\ \widehat{\beta_{IV}} = \frac{\beta cov(z_i, x_i)}{cov(z_i, x_i)} + \frac{cov(z_i, \varepsilon_i)}{cov(z_i, x_i)}. \qquad 6$$

From equation 6, the second assumption of IV becomes evident. The second assumption is the relevance assumption, or that the instrument must be statistically associated with the endogenous variable (x_i). As can be seen in equation 6, this means, in mathematical terms, $cov(z_i, x_i) \neq 0$. Covariance does not imply a direction to the relationship, whether AOD (our instrument) determines deforestation or deforestation determines AOD (or neither) is irrelevant, as it is the covariance between the two that is important.

By these two assumptions of IV, that $cov(z_i, \varepsilon_i) = 0$ and $cov(z_i, x_i) \neq 0$, equation 6 simplifies to $plim\ \widehat{\beta_{IV}} = \beta$, illustrating IV is a consistent estimator of the true relationship.

Weak Instruments: Equation 6 also illustrates another important aspect when considering the application of instrumental variables, and that is a problem known as "weak instruments." The problem occurs if the exclusion criteria, $cov(z_i, \varepsilon_i) = 0$, fails. Based on the relationship between covariance and correlation (namely, $cov(x, y) = corr(x, y) * \sigma_x \sigma_y$ where σ is the standard deviation of each variable) and assuming $cov(z_i, x_i) \neq 0$, we can rewrite equation 6 to illustrate the problem (omitting subscripts for simplicity).

$$plim\ \widehat{\beta_{IV}} = \beta + \frac{corr(z, \varepsilon) * \sigma_z \sigma_\varepsilon}{corr(z, x) * \sigma_z \sigma_x} = \beta + \frac{corr(z, \varepsilon) * \sigma_\varepsilon}{corr(z, x) * \sigma_x}. \qquad 7$$

If there is a small correlation between the instrument and the error, the last term in equation 7 does not drop out and the IV estimator is inconsistent ($plim\ \widehat{\beta_{IV}} \neq \beta$). If $corr(z, \varepsilon)$ is just slightly different from zero and $corr(z, x)$ is much different than zero, the last term is of minimal influence. However, if the instrument is only weakly correlated with the endogenous covariate, the last term of equation 7 can become large. In practice, weak instruments can cause the IV estimator to be severely biased. Since there is no test to validate the exclusion criteria, the strength of the relationship between the instrument and the endogenous variable is very important in practice, and can be formally tested[25] as in the supplementary material from MacDonald and Mordecai.[4]

Conclusion: Understanding the effects of environmental change on infectious disease transmission—from diseases long endemic to the tropics like malaria, to novel emerging pathogens we have yet to discover like SARS-COV-2—is of fundamental and increasing importance. In these complex socio-ecological systems that are difficult to study experimentally, emerging data sources (e.g., high spatio-temporal resolution earth observation data) and causal inference methods (e.g., IV estimation) represent one methodological approach that can help us achieve such clearer understanding.

Acknowledgements: We would like to acknowledge Dr. Ashley Larsen and two anonymous reviewers for their thoughtful comments and feedback on this manuscript.

Financial Support: AJM and EAM were supported by the National Science Foundation and the Fogarty International Center (DEB-2011147). EAM was supported by the National Science

Foundation (DEB-1518681), the National Institute of General Medical Sciences (R35GM133439), the Stanford King Center for Global Development and the Terman Award.

Conflicts of Interest: The authors declare no conflicts of interest.

Author Contact Information:

Andrew J. MacDonald: Bren School of Environmental Science and Management, University of California, Santa Barbara, CA 93106-5131; andy.j.macdon@gmail.com

Erin A. Mordecai: Department of Biology, Stanford University, Stanford, CA 94305; emordeca@stanford.edu

References

1. Plowright RK, Reaser JK, Locke H, Woodley SJ, Patz JA, Becker DJ, Oppler G, Hudson PJ, Tabor GM, 2021. Land use-induced spillover: a call to action to safeguard environmental, animal, and human health. *Lancet Planet Health* 5(4):e237-e245. doi:10.1016/S2542-5196(21)00031-0.

2. Thomas MB, 2020. Epidemics on the move: Climate change and infectious disease. *PLoS Biol* 18(11):e3001013–2. doi:10.1371/journal.pbio.3001013.

3. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P, 2008. Global trends in emerging infectious diseases. *Nature* 451(7181):990-993.

4. MacDonald AJ, Mordecai EA, 2019. Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing. *Proc Natl Acad Sci USA* 116(44):22212-22218. doi:10.1073/pnas.2014828117.

5. Valle D, Laporta GZ, 2021. A Cautionary Tale Regarding the Use of Causal Inference to Study How Environmental Change Influences Tropical Diseases. *Am J Trop Med Hyg*. doi:10.4269/ajtmh.20-1176.

6. Larsen AE, Meng K, Kendall BE, 2019. Causal analysis in control–impact ecological studies with observational data. *Methods Ecol Evol* 10(7):924-934. doi:10.1111/2041-210X.13190.

7. Bonds MH, Dobson AP, Keenan DC, 2012. Disease Ecology, Biodiversity, and the Latitudinal Gradient in Income. *PLoS Biol* 10(12):e1001456. doi:10.1371/journal.pbio.1001456.

8. MacDonald AJ, Larsen AE, Plantinga AJ, 2019. Missing the people for the trees: Identifying coupled natural–human system feedbacks driving the ecology of Lyme disease. *J Appl Ecol* 56(2):354-364. doi:10.1111/1365-2664.13289.

9. Bauhoff S, Busch J, 2020. Does deforestation increase malaria prevalence? Evidence from satellite data and health surveys. *World Dev* 127:104734. doi:10.1016/j.worlddev.2019.104734.

10. Jones IJ, et al., 2020. Improving rural health care reduces illegal logging and conserves carbon in a tropical forest. *Proc Natl Acad Sci USA* 117(45):28515-28524.

11. Garg T, 2019. Ecosystems and human health: The local benefits of forest cover in Indonesia. *J Environ Econ Manage* 98(24):102271. doi:10.1016/j.jeem.2019.102271.

12. Couper LI, MacDonald AJ, Mordecai EA, 2021. Impact of prior and projected climate change on US Lyme disease incidence. *Glob Chang Biol* 27(4):738-754. doi:10.1111/gcb.15435.

13. Larsen AE, MacDonald AJ, Plantinga AJ, 2014. Lyme Disease Risk Influences Human Settlement in the Wildland-Urban Interface: Evidence from a Longitudinal Analysis of Counties in the Northeastern United States. *Am J Trop Med Hyg* 91(4):747-755. doi:10.4269/ajtmh.14-0181.

14. Santos AS, Almeida AN, 2018. The Impact of Deforestation on Malaria Infections in the Brazilian Amazon. *Ecol Econ* 154:247-256.

15. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. first edition. Cambridge, Massachusetts: MIT Press; 2002.

16. Morgan WT, Darbyshire E, Spracklen DV, Artaxo P, Coe H, 2019. Non-deforestation drivers of fires are increasingly important sources of aerosol and carbon dioxide emissions across Amazonia. *Sci Rep* 9:16975. doi:10.1038/s41598-019-53112-6.

17. Aragão LEOC, et al., 2018. 21st Century drought-related fires counteract the decline of Amazon deforestation carbon emissions. *Nat Commun* 9:536. doi:10.1038/s41467-017-02771-y.

18. Chen Y, Morton DC, Jin Y, Collatz G, Kasibhatla PS, van der Werf GR, DeFries RS, Randerson J, 2013. Long-term trends and interannual variability of forest, savanna and agricultural fires in South America. *Carbon Manag* 4(6):617-638. doi:10.4155/cmt.13.61.

19. Hengl T, Wheeler I, 2018. Soil organic carbon content in x 5g / kg at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250m resolution (Version v0.2). doi:10.5281/zenodo.2525553.

20. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R, 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens Environ* 202(C):18-27. doi:10.1016/j.rse.2017.06.031.

21. Mordecai EA, et al., 2012. Optimal temperature for malaria transmission is dramatically lower than previously predicted. *Ecol Lett* 16(1):22-30. doi:10.1111/ele.12015.

22. Burke M, Hsiang SM, Miguel E, 2015. Global non-linear effect of temperature on economic production. *Nature* 527(7577):235-239. doi:10.1038/nature15725.

23. Cattelan AJ, Dall'Agnol A, 2018. The rapid soybean growth in Brazil. *OCL* 25(1):D102. doi:10.1051/ocl/2017058.

24. Viana JS, Gonçalves EP, Silva AC, Matos VP, 2013. *Climatic Conditions and Production of Soybean in Northeastern Brazil*. IntechOpen. doi:10.5772/52184.

25. Olea JLM, Pflueger C, 2013. A robust test for weak instruments. *J Bus Econ Stat* 31(3):358-369. doi:10.1080/00401706.2013.806694.