Inter-model comparison of the landscape determinants of vector-borne disease: implications for epidemiological and entomological risk modeling

Alyson Lorenz¹, Radhika Dhingra¹, Howard H. Chang², Donal Bisanzio³, Yang Liu¹, Justin V. Remais¹,⁴*

¹ Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA.
² Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, USA.
³ Department of Environmental Sciences, Emory University, Atlanta, GA, USA.
⁴ Program in Population Biology, Ecology and Evolution, Graduate Division of Biological and Biomedical Sciences, Emory University, Atlanta, GA, USA.

*Email: justin.remais@emory.edu
Development of a checklist for the evaluation of models for extrapolation

Checklists have been successfully used in a variety of disciplines in order to improve the quality and reliability of work. In both public health and ecology, checklists have been used to organize a broad literature (e.g., Bosshard 2000; de Groot et al. 2002; de Groot et al. 2003), to assess the quality of epidemiologic studies, (e.g., Downs and Black 1998) or ecological management strategies (e.g., Lindenmayer et al. 2008), and to facilitate communication and comparison of related work (e.g., Bosshard 2000; Vandenbroucke et al. 2007). To improve the methodological quality of landscape model extrapolation in vector-borne disease risk assessment, we provide checklist criteria for the qualitative assessment of models being considered for extrapolation. Here, we briefly describe model characteristics that appear on the checklist, and distinguish between conceptual extrapolation and spatial extrapolation.

Conceptual extrapolation

Applying existing models to new research questions—a conceptual extrapolation—raises the importance of key characteristics of the original analysis, including its predictor and outcome variables, scale and resolution, reproducibility, and data quality and availability.

Characteristics of the original analysis. Ascertaining the quality of the original analysis includes assessment of the modeling technique used, the species being modeled,
and the model selection technique employed. For species distribution modeling, multi-
model comparisons have found that different modeling techniques (e.g., generalized
linear models, artificial neural networks, etc.) exhibit varying generalizability,
robustness and predictive capacity (Pearson et al. 2006; Randin et al. 2006; Dormann et
al. 2008), and thus no one class of model is universally preferred for conceptual
extrapolation (Dormann et al. 2008; Jeschke and Strayer 2008; Elith and Leathwick
2009). Models drawn from research conducted on the exact species of interest are
preferred (MacArthur 1958; Elith and Leathwick 2009); however, at times, a model
describing a related organism may be useful (Raxworthy et al. 2003). The model
selection technique employed in the original analysis should be examined, as models
selected using fit as the sole criterion may not be appropriate given that over-fit models
are often uninformative beyond the spatial or temporal confines of the original analysis
(Ginzburg and Jensen 2004; Hitchcock and Sober 2004). Additional criteria such as
parsimony, agreement with previous findings or theories, and predictive capability can
strengthen the model selection process (Ginzburg and Jensen 2004).

Predictor and outcome variables. The predictor and outcome variables that appear
in a model, and how they were selected and treated in the analysis, are primary
considerations for determining the model’s value and generalizability. Using indirect
predictor variables (e.g., elevation), which act as proxies for other influential variables
(e.g., temperature, humidity), may decrease predictive ability (Austin and Smith 1989; Elith and Leathwick 2009). Proxy variables, often chosen because of limited data availability (Dormann et al. 2008), should be selected only after a careful assessment of the appropriateness of the proxy in light of the underlying variables of interest. The data type (continuous, categorical, nominal, ordinal, etc.) utilized in the original analysis may affect the utility of a candidate model, as inaccurate assumptions about the true distribution the data (Sauerbrei et al. 2007), and inappropriate categorical transformations (Brooker et al. 2002; Araujo et al. 2005; Royston et al. 2006), can contribute to bias yielding a model unsuitable for conceptual extrapolation.

*Scale and resolution.* At different spatial and temporal scales and resolutions of analysis, both the magnitude and direction of a relationship between predictor and outcome variables can vary (Chase and Leibold 2002; Jackson et al. 2006). In the spatial domain, such variation may be produced by disappearance of non-dominant habitat types with decreasing resolution, which influences variable collinearity, or bias associated with alternate methods of averaging over areas (Turner et al. 1989; Benson and MacKenzie 1995). Models which were developed using data at coarse resolution may have low predictive ability if used at fine scale (Kunin 1998), for instance. Thus, a model fit at a scale that is mismatched with the scale of the new research question may be inappropriate for conceptual extrapolation. Likewise, the predictive ability of species
distribution models across timescales may also be limited if, for instance, species rapidly
 evolve leading to changes in habitat suitability and violating the assumption of niche
 conservatism (Pearson and Dawson 2003; Chaves and Koenraadt 2010).

Reproducibility. Logistical issues in reproducing candidate models also arise. When
 researchers fail to specify the full model, or do not completely define the methods used
 for data transformation or for dealing with missing data, it may be impossible to re-
 apply the model unless the original authors can be consulted. Such unspecified models
 were excluded from our analyses. Access to appropriate computer applications and
 versions may also be a barrier to successful replication of the model.

Data quality and availability. Finally, the quality of the data used to fit candidate
 models is an important consideration when choosing a model for conceptual
 extrapolation (Dormann et al. 2008). In many cases, the most robust parameter
 estimates result from species distribution models fit to datasets that balance quantity
 with quality, i.e., achieving a sufficient number of points with adequate quality
 (Dormann et al. 2008).

Spatial extrapolation

Extrapolation of models across spatial domains beyond what was included in the
 original analysis requires many of the same considerations as when undertaking
 conceptual extrapolation, but must also take into account additional issues related to
the spatial aspects of model variables, choice of predictor and outcome variables, and spatial extent, as described briefly here.

**Spatial aspects of model variables.** Issues surrounding spatial extrapolation of species distribution models are covered in a wide body of literature that has been comprehensively reviewed elsewhere (e.g., Miller *et al.* 2004; Peters *et al.* 2004). Several key issues routinely arise, such as the fact that suitable habitats tend to be more varied in the center of the geographical range occupied by a species, leading the relationships between predictors and outcomes to differ by range position (Peterson *et al.* 2000; Swihart *et al.* 2003; Randin *et al.* 2006). Thus applying a model fit at the center of a species’ range to an area at the edge of the range could lead to overestimation of presence at the edge, for instance, or underestimation of presence at the center. Models must therefore incorporate locations throughout a species’ range (e.g. Webber *et al.* 2011), and testing models against independent observations rather than observations used in model fitting is of course preferred (Fielding and Bell 1997; Guisan and Zimmerman 2000).

**Choice of predictor and outcome variables.** The relevance, range, and relationships between predictor and outcome variables often differ across space. Issues of spatial stationarity of predictor-outcome relationships are well-described elsewhere (e.g., Brunsdon *et al.* 1998). Predictor variables may become irrelevant in new geographical
areas where these variables are missing or where their explanatory power decreases (Rodder and Lotters 2010). The use of indirect or proxy predictors, which may fail to describe true habitat preferences of a species, may also limit a model’s spatial transferability (Austin and Smith 1989; Austin 2002). Even when the exact predictor of interest can be measured across the extrapolation zone, its numerical range may fall outside the range over which the original model was fit, limiting model performance (Peters et al. 2004). Finally, the categorization of outcome variables may have limited meaning in new locations. For example, the category of ‘high’ risk for mosquito bites may represent a different range of risk values in Albany, New York than in New Orleans, Louisiana.

Spatial extent. The spatial extent of the data used to fit the model, and that of the new domain, raise similar issues. Research conducted over a large area in which there is significant variance in predictors and outcomes may lead to a more robust and transferable model, particularly with respect to extreme values (Wiens 1989). Additionally, predictor data must be available across the extrapolation zone, and attention must be paid to missing data and data collected at a resolution that differs from the intended resolution of the analysis (Kistemann et al. 2002).
References


