Maps and models are never valid, but they can be evaluated

D G Rossiter

david.rossiter@wur.nl; d.g.rossiter@cornell.edu

At Pedometrics 2017 I began my invited talk on "Past, present, & future of information technology in pedometrics" with a slide explaining why I think pedometricians should not use the term "model validation", but rather call this process "model evaluation". That slide provoked quite some comment, so I thought it would be useful to explain the reasoning for this proposed semantic shift.

This is by no means my idea. It originated more than 20 years ago in two seminal articles written by the historian of science Naomi Oreskes (1994, 1998)[1], presently Professor of the History of Science at Harvard University; these are well worth reading to get a deeper philosophical and practical justification than I give here. I was introduced to these articles and the use of "model evaluation" about five years ago by colleague Janneke Ettema at the University of Twente, an atmospheric scientist faced, like pedometricians, with modelling natural systems. I was convinced and have tried ever since to refer to "model evaluation". As I will explain, this is not only more correct, especially in communication with decision makers, but also opens up possibilities for deeper discussions of our models than are possible with the term "model validation".

Oreskes' context was the application of numerical simulation models of natural systems for predictions, these to be used to direct public policy, and the communication to decision makers of the uncertainty in the information obtained from these models. Pedometrics is active in this sort of modeling, see for example the Vereecken et al. (2016) review of soil process modelling; (these authors consistently use the term "validation"). However, we are rarely directly in the line of fire of the public and decision makers, especially in our papers written for our peers. Still, our soil maps are used for contentious issues such as soil pollution, wetland regulation, and agricultural subsidies. Our pedotransfer functions are used in hydrologic models for flood hazard mapping and surface and groundwater pollution maps. Our models for soil hydrology, soil survey interpretations and land valuation have legal force in some contexts.

The core of Oreskes' sophisticated argument, as I rephrase it, is that the word "valid" in English and any language where this word derives from the Latin "validus" (strong, powerful, effective), the relevant definition in the OED is: "Of arguments, proofs, assertions, etc.: Well founded and fully applicable to the particular matter or circumstances; sound and to the point; against which no objection can fairly be brought." In normal speech we use phrases such as "valid argument", meaning that it is correct. A related meaning is used in "valid passport", i.e., the passport can be used to travel between political entities. Thus, the usual sense of "valid" is "true, not false", a binary concept. If I take my birth certificate to the local administration to establish my identity, they require that it be "validated" by the responsible official where it was issued. It is then a "valid", and can be used to establish my identity. Oreskes: "[validity] denotes the establishment of legitimacy, typically given in terms of contracts, arguments, and methods." In pedometrics we can speak of a "valid" method for a laboratory determination or sampling design, but only in the sense that the method has been established as a standard.

---

[1] Naomi Oreskes has written several semi-technical and popular books on the history of science and science policy, well worth a place on your nightstand. I especially enjoyed her edited compilation of first-person accounts on how the theory of plate tectonics was proposed, rejected, confirmed and accepted: *Oreskes, N. (Ed.). (2001). Plate tectonics: an insider's history of the modern theory of the Earth. Boulder, Colo.: Westview Press*. This story is told much more briefly in *Oreskes, N. (2013). Earth science: How plate tectonics clicked. Nature News, 501(7465), 27. DOI: 10.1038/501027a*. She has also dealt with how unscrupulous scientists may play-for-pay: Oreskes, N., & Conway, E. M. (2010). Merchants of doubt. New York: Bloomsbury Press.

Maps and models are never valid, but they can be evaluated

The public, including decision makers, thus equate talk of a "valid" model with one that is correct and true. What is at stake is scientific credibility. The term "valid" is only positive, and the public is rightly suspicious of scientists who over-sell their models either because these scientists cynically feel the public does not understand nuances, or because they consider that the policy is too important to be left to the public, or because they ignore Niels Bohr (quoting Robert Storm Petersen, Danish cartoonist): "Prediction is very difficult, especially if it's about the future". Oreskes provides a devastating analysis of all three faults in the infamous Club of Rome "The Limits to Growth" models from the early 1970s. The spectacular failure of their predictions set back serious discussions of what are in fact the limits to growth.

In pedometrics or any kind of statistical, process, or simulation modelling we never claim that the results are valid in this sense. Indeed, that is why we compute so-called "validation statistics" and hold out "validation samples" on which we compute these. But the result of these computations is some "degree of validity", which is a contradiction in terms. The statement that the RMSE, RPD etc. have certain values when the model is applied to independent observations is certainly useful information, but even an RMSE of zero does not prove a model is valid in the wider sense, and as the public uses the term. That is, the model may not be a correct simplified representation of reality, even if the match with observations is good. The classic example here is the Ptolemeic model of the heavens, which, once calibrated, had excellent agreement with observations as known prior to 1610 but failed when Galileo showed the impossibility of the model to explain the phases of Venus.

Pedometric models and their outputs have many sources of uncertainty. Some sources are more or less easily quantified, for example, uncertainty of calibration data due to sampling error and measurement procedures. We can account for these by simulation and sensitivity analysis. Uncertainty in model form can be quantified by comparing the results of various plausible model forms on the same data. These are good examples of quantitative evaluation. But some uncertainty cannot be measured; in particular, are all relevant processes or factors included in the model? This is of primary importance for evaluation of predictive power in new contexts, which are especially of interest in public policy discussions. A recent and controversial example are models of possible C sequestration in soils used to promote the "4 pour mille" concept (Minasny et al. 2017).

What are we really doing in the so-called "validation" step of modeling? Answer: we are evaluating the model. That is, we are determining to what degree the model is useful in our problem, whether the model form and assumptions are justified, whether the data at hand are sufficient to give a reliable answer, how far the model can be extrapolated in space and time. The OED sense of "evaluation" here is "the action ... estimating the force of probabilities, evidence, etc.". Another sense "the action of appraising or valuing", in our case appraising the value of the model to our problem. For example, how useful is a predictive map made by digital soil mapping techniques? What level of certainty can the map user expect? Is this sufficient for the map user's proposed applications?

On the positive side, the term "evaluation" opens up the discussion to more than statistical measures of output agreement with independent observations. We can now discuss the model form, model assumptions, modelling choices, selection of evaluation criteria, selection of evaluation sample etc. -- i.e., we are "evaluating" the entire modeling process, a much richer implication than the word "validating" can provide. Oreskes: "Quality can be evaluated in several ways: on the basis of the underlying scientific principles, on the basis of the quantity and quality of input parameters, and on the ability of the model to reproduce independent empirical data. All of these things can be discussed, but none of them should be discussed in either/or terms." So what we now refer to as "validation statistics" can be better described as "agreement of model output with independent observations". This then forces us to describe the plan by which independent observations were selected, and what population they represent.

Oreskes identifies model flaws of four kinds: theoretical, empirical (imprecise or limited measurements), parametrical and temporal. All of these can lead to disagreement of model output with reality. But which are causing this? Theoretical flaws are due to our poor understanding of processes, and these are quite difficult to identify. The empirical flaw is of course dominant in pedometric studies -- our observations are a tiny fraction of the population. Parametrical flaws arise with model simplification, e.g., assuming isotropy and second-order stationarity when computing an empirical variogram from limited observations. Temporal flaws, which pedometricians might revise to spatio-temporal flaws, arise when extrapolating into the future or unobserved regions. There is no way to know that conditions will be the same as those that produced the observations on which the calibrated model is based.

Maps and models are never valid, but they can be evaluated

Consider ordinary kriging. Theoretical flaws refer to the theory of random fields on which kriging is based -- is this an accurate representation of local soil spatial variability? Is the random field second-order stationary? Further, even supposing we have such a random field, does the selected variogram model form correctly represent its structure? Empirical flaws are due to the sparse sample on which we calibrate a variogram model. Even given a proper variogram model, the parametrical flaw is the fitting of variogram parameters from the limited sample. Finally, the spatio-temporal flaw is in assuming the fitted model can be extrapolated. This does not imply that maps made with ordinary kriging are not useful. The so-called "validation statistics" do give some idea of predictive accuracy within the context in which the original study was done, and if these are satisfactory we have some basis for using the resulting map. This is a reasonable "evaluation" criterion for the model.

What then do we call the various terms that have been traditionally used in pedometrics papers?

1. "Model validation" becomes "model evaluation", and should be described in the broader sense explained above, as "fitness for use" and “appropriateness of modelling approach”.

2. "Validation dataset" (or "observations") becomes "independent observations [not used in modelling]". Another possibility is "Assessment dataset" (or "observations"). This is a short way to contrast this with "calibration dataset": Calibration vs. (quantitative) assessment.

3. The process of numerically comparing model predictions to independent observations is one part of model evaluation, and should be referred to as "predictive accuracy assessment". This supposes that the assessment dataset is representative of the target population. Note that this need not be the population sampled for calibration, if the interest is in extrapolation.

4. "Validation statistics" becomes "agreement between model and observations".

5. The term "cross-validation" is retained, because (1) "cross-evaluation" sounds awkward, (2) it is limited to a specific method of using the data to evaluate the model, and (4) is rarely used in communication with decision-makers. However, the terms "to cross-validate" or "was cross-validated" should not be used, instead something like "model output and observations were compared with statistics from 10-fold cross-validation ..."

I illustrate these proposals with some revisions of phrases from a recent paper on which I am co-author (Zeng et al. 2016):
Original: "First, due to the low density of samples in this study area, we used leave-one-out cross-validation to evaluate the results"
Revised: "First, due to the low density of samples in this study area, we used leave-one-out cross-validation to evaluate the predictive accuracy of the results"

Original: "...each sample was validated individually on the basis of the calibration set compiled from the remaining dataset"
Revised: "...each observation was compared to its prediction made from the model calibrated on the basis of the other observations"

Original: "*n* is the total number of validation samples"
Revised: "*n* is the total number of observations not used in model calibration"

Original: "...the leave-one-out cross validation method was used to validate the mapping results."
Revised: "...the leave-one-out cross validation method was used to assess the relative accuracy of the mapping results."

Evaluations of this proposal are welcome!

References:

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. Science, 263, 641--646. https://doi.org/10.1126/science.263.5147.641

Oreskes, N. (1998). Evaluation (not validation) of quantitative models. Environmental Health Perspectives, 106(Suppl 6), 1453--1460.

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., & Winowiecki, L. (2017). Soil carbon 4 per mille. Geoderma, 292, 59-86. https://doi.org/10.1016/j.geoderma.2017.01.002

Vereecken, H., Schnepf, A., Hopmans, J. W., Javaux, M., Or, D., Roose, T., & Young, I. M. (2016). Modeling Soil Processes: Review, Key Challenges, and New Perspectives. Vadose Zone Journal, 15(5). https://doi.org/10.2136/vzj2015.09.0131

Zeng, C., Zhu, A.-X., Liu, F., Yang, L., Rossiter, D. G., Liu, J., & Wang, D. (2017). The impact of rainfall magnitude on the performance of digital soil mapping over low-relief areas using a land surface dynamic feedback method. Ecological Indicators, 72, 297-309. https://doi.org/10.1016/j.ecolind.2016.08.023

Careful writing should make that clear, adoption of a half-baked linguistic rule will not

Murray Lark

I am entirely unconvinced by the argument that David has put forward, and I think it quite wrong. First, though, I do not think we are in any fundamental disagreement about what the process of model validation does and does not do. Tom Addiscott, formerly leader of the soil process modellers at Rothamsted, strongly advocated the use of the word “validate” to describe the process of testing model predictions against data, as opposed to “verify”. He argued this on the basis that “verify” means to establish truth, whereas “validate” means to establish validity which means “strength, or fitness for purpose”.

Why is this an important distinction? In modern English the word “true” can be affirmed or denied only in respect of a proposition, some statement of fact. For example “there is no largest prime number”, “there will be earthquakes in East Acton tomorrow”, or “the pH of the topsoil at location x is 6.5”. These propositions can be verified, the first one by mathematical proof, and the second two by empirical observation. If I visit x and measure the soil pH and get a value of 6.2, then the prediction was false. However, the pedometrician would not conclude that the prediction was meaningless on this basis. She knows that any prediction has attendant uncertainty, and her assessment of it will take account of this, comparing, for example, the magnitude of the prediction error with observed natural or analytical variation. She might also observe whether the prediction, while not precise, is useful to the manager. Will it lead to a bad decision? In short, she assesses the usefulness of the prediction, and the evidence that it encodes information about the variable of interest. That is what we mean by its “strength, applicability, well-foundedness”, which are synonyms of “validity” in the Shorter Oxford English Dictionary’s definition of “valid”.

Now David’s argument, in my opinion, takes a wrong turn where, having observed that the English “valid” derives from Latin "validus" (strong, powerful, effective), he segues into the quite unfounded statement that

In normal speech we use phrases such as "valid argument", meaning that it is correct.

Well I must disagree. A valid argument is one that is correctly constructed. It is a formal property of the argument. If someone offers an argument, the conclusion of which is certainly true, it does not follow that the argument is valid. For example

Some bears are brown
Grizzlies are bears
Grizzlies are brown

is not a valid argument, although it has a true conclusion and true premises. We can test the conclusion in the field, but the validity of the argument is tested by the logician. In modern logic we turn the argument into symbols and manipulate these to test validity. This process would show that the structure of the argument is not sound. Consider another argument

All bears are brown
Polar-bears are bears
Polar-bears are brown

This is a valid argument, but its conclusion is untrue. The classical logician would identify the argument as an example of Aristotle’s valid syllogism in BARBARA, if the premises are correct then the conclusion must be. The zoologist notes that the first premise is wrong, which is why the valid argument can lead to a false conclusion, as it does in this case.

So I reject the idea that, in English, a “valid argument” means an argument with a true conclusion. It does not mean that at all. Rather it means an argument which is sound and fit for purpose. Exactly what we want to establish about a method for pedometrical prediction.

David goes on to say

> *A related meaning is used in "valid passport", i.e., the passport can be used to travel between political entities. Thus, the usual sense of "valid" is "true, not false", a binary concept.*

I think this argument a complete non-sequitur. A passport is valid if it has not been withdrawn by the issuing authority, if it has not expired, if it has not been defaced etc. That makes it fit for purpose, “valid” in the sense of The Oxford English Dictionary, and Tom Addiscott’s “model validation”. We cannot say that a passport is “true” or “false”. What about a forged passport? Well here again a bit of logic, in particular Bertrand Russell’s logic, will help us. Grammatically the expressions “Dutch Passport”, “Blue Passport” and “Forged Passport” are the same, an adjective qualifies a noun. In the first two cases the adjective tells us what kind of passport we are dealing with. In the third case it actually tells us that the object in question is not a passport (a document issued by a competent authority), but is rather a document run up by some shady character in exchange for a handful of used ten pound notes. Russell would have recognized this as an example of a case in which the structure of ordinary language obscures the logical structure of an expression, leading us into muddles. I respectfully suggest that David has ended up in such a muddle, because the only sense that can be given to the expression “valid passport”, is “fit for purpose”, back to Tom Addiscott and the OED.

When we have a pedometrical model we have a statistical prediction, a conditional expectation (or some other moment of the conditional distribution of the variable of interest) and one or more measurements of the uncertainty of that quantity, treated as a prediction of an unknown variable. When we validate that model we check whether the prediction, as compared with the observation, gives us grounds to regard the model as fit for purpose, and “doing what it says on the tin”. There are various statistics which we might use to help with this, and some of them, pace David’s argument, will indeed indicate degrees of validity. A prediction set with an RMSE of 1.0 comes from a model with a greater degree of validity (strength, fitness for purpose) than a prediction set with an RMSE of 10, although both may be useful if the standard deviation of the variable in question is 100. The standardized square prediction error, or coverage probabilities of the prediction, tell us whether the quantification of the uncertainty in the predictions seems to be sound (which might not be the case if, for example, the variogram has been influenced by outlying data).

The Shorter Oxford English Dictionary gives a particular sense of the verb “to validate”. It reads thus

> *[in] Computing etc. …. confirm or test the suitability of (a system, program, etc.)*

I cannot think of a better way to describe what we do in pedometrics when we validate predictions; and I, for one, will fight to preserve such a valuable word.

I very much hope that editors of soil science journals and others will not take David’s recommendation on board. They should, however, insist that authors are always clear and explicit about what they are doing when they validate predictions, and the claims that are based on such validations. We are not claiming to show that predictions are “true” (even where careful analysis could attach meaning to such a statement), we are testing the system used to generate the predictions and quantifying their fitness for purpose. Careful writing should make that clear, adoption of a half-baked linguistic rule will not.

We can avoid confusion and misunderstanding if we make the change

Gerard Heuvelink

Agreement on the meaning of terms is crucial to science. If different people interpret terms differently then this may cause confusion and obstruct scientific progress. One of the terms that is regularly debated is ‘validation’, because its meaning is not the same to all. For instance, in my experience soil physicists and soil hydrologists interpret “the model has been validated” as “the model has been proven suitable”, while my interpretation and that of most pedometricians of “the model has been validated” is “the model predictions were put to the test and compared with independent observations”. So even if a model does a very poor job, has a large Mean Squared Error and low Amount of Variance Explained, it would still be considered validated.

David reminds us of the work of Naomi Oreskes that defines validation as “the establishment of legitimacy” and states that a valid model “does not contain known or detectable flaws and is internally consistent”. This definition agrees fairly well with the interpretation of soil physicists and soil hydrologists and refutes that of us, pedometricians. Even closer to the interpretation of soil physicists and soil hydrologists comes the definition of Rykiel (1996), who defines model validation as “a demonstration that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model”.

So should we abandon the term ‘validation’ and replace it with ‘evaluation’ when we refer to the process of comparing model predictions with independent observations? I started writing this comment having in mind that my answer to this question is ‘no’ (because replacing validation with evaluation is not a solution in all cases, because making the change in practice may prove to be very difficult, and because we pedometricians know very well what we mean by validation so why bother?), but the more I think about it the more I agree with David. It looks as if our interpretation of model validation is quite different from that used in mainstream science. This is clearly undesirable because we do not live on an island and need to communicate well with other (soil) scientists. We can avoid confusion and misunderstanding if we make the change.

So David has my full support. From now on I will try to use evaluation instead of validation whenever appropriate, although I realise that this may be difficult because it is oh so easy to fall back into old habits!

Gerard Heuvelink

Reference

Rykiel, E.J. (1996), Testing ecological models: the meaning of validation. Ecological Modelling 90, 229-244.

A problem well-stated is a problem half solved

Philippe Lagacherie

David Rossiter provides us with a bright and very convincing argument for replacing “validation“ by “evaluation” in the DSM papers. This could be seen as some kind of a Don Quichotte’s enterprise since everybody in the DSM community has used validation for a long time now and it is never easy to change habits, especially the bad ones. Yet, it is far to be a vain fight. Remember Kettering’s quote saying “a problem well-stated is a problem half solved”. Furthermore, as David points out, “validation” is misleading for users of our soil maps which are not familiar with our jargon. Therefore, I thank David for rowing upstream and I invite my colleagues rowing with him.

That being said, I would like to take an opportunity to stress that model validation – or model evaluation – remains in 2017 a third-class passenger - if not the clandestine one - of our way towards an operational DSM. Although valuable contributions have been brought in the past to set the theoretical framework of model evaluation, we are still lacking in ‘valid’ and realistic solutions that could be really applied on real study areas. The current DSM literature does not help us a lot in dealing with some basic questions like determining how large an ‘assessment dataset’ must be, deciding whether a difference of 0.02 in R^2 means a difference in performances between two models or not, or evaluating how much the uncertainty is underestimated by the often-unescapable cross-validation method. A more holistic approach of model evaluation that would include knowledges on expected soil patterns and on limitations of

A valid map is a map that is made by a valid model

Dick Brus

Introduction

First of all I would like to congratulate David Rossiter with his nicely written, thought provoking article on map validation and evaluation. The article forced me to gather my thoughts. This comment is the result of this struggle, I hope it is of use for others .

Model validation

Validation is used both for models and for maps: model validation and map validation. A map can also be seen as a model, but here I use model for a deterministic or statistical equation. Let me start with model validation, more specific validation of statistical models. In my basic statistics classes I teach the students that in linear regression modelling several assumptions are made:

- a linear relation between response variable and independent variables (predictors)
- the data are independent
- the variance is constant
- if we want confidence intervals of estimated means and or prediction intervals for unobserved units in the population, we also assume a normal distribution for the residuals

I stress that we always should check the validity of these assumptions, for instance by making scatter plots of residuals against fitted values to check the linearity and the constant variance assumptions, making a Q-Q plot of the residuals to check the assumption of a normal distribution, and estimating an experimental variogram of the residuals to check the assumption of independent data. I also point them to formal test such as the Moran’s I test for spatial autocorrelation of the residulas, the Shapiro-Wilk test for normality and the Levene’s test for constant variance within groups. If none of the modelling assumptions is clearly violated, we treat the model as a valid model. It is like statistical testing of an hypothesis, the null hypothesis being that the model is valid. Unless there is clear evidence or theory against the null hypothesis, it is not rejected, and we take the model as a valid model.

Map validation

If there is no evidence that one or more assumptions are violated, we can use the model to predict for unobserved units in the population, i.e. mapping. What is usually done in map validation is computing map quality indices such as the mean error (ME), mean squared error (MSE), standardized squared prediction error (SSE) *et cetera* for continuous soil maps, and the overall, user’s and producer’s accuracies for categorical maps. The closer the map quality indices are to their ideal values, the more this supports the validity of the model. So when the aim of estimating the map quality indices is to check the validity of the model underlying the map, these quality indices serve as validation statistics, and the estimation of these statistics can be named *model* validation. An accuracy plot (Goovaerts, 2001) and a variogram of the prediction errors are examples of graphs that are also tailored at checking the validity of the modelling assumptions. If we define a valid *map* as a map that has been made by a valid *model*, the estimation of map quality indices and graphs *tailored at checking the validity of the model* can also be named *map* validation.

David Rossiter states that ‘even an RMSE of zero does not prove the model is valid’. David illustrates the problem with the nice example of the Ptolemeic model of the theory of the heavens. Even if the match between predictions and observations is perfect, this does not prove the validity of the model. This is entirely analogous to statistical hypothesis testing. The null hypothesis (in our case ‘the model is valid’) simply cannot be proven. We can only reject it, so that we conclude that it is very unlikely that the model is valid (very likely that it is invalid), or not reject it. In the latter case we have not proven that the

model is valid, we only have not enough evidence against it. Making this clear does not convince me that the estimation of map quality indices such as the RMSE cannot be tailored at map validation, and therefore should not be named map validation, but preferably map evaluation.

Valid estimates of map quality indices

I would like to stress here that the map quality indices ME, MSE, SSE, overall, user's and producer's accuracies et cetera should be defined in terms of population parameters, not as sample averages. For instance, the ME should be defined as the average of the errors over all N units in the population. A subset of these units (sample) is used to estimate this population ME. We are uncertain about the population ME, and it is important to quantify our uncertainty, for instance by estimating its standard error. We can then statistically test the null hypothesis 'the population ME equals zero'. If this hypothesis is not rejected, this supports the correctness, validity of the model underlying the map; if it is rejected than we have evidence that the model and map is not valid.

As argued in our 'Sampling for validation' paper (Brus et al., 2011) the validation sample can best be selected by additional probability sampling, so that the map quality indices can be estimated by design-based inference. Design-based estimates are model-free, no modelling assumptions are made, and as a consequence no objections can be made against the estimated map quality indices and their standard errors. In other words, with a design-based sampling strategy we obtain valid estimates of the map quality indices. The quality of these estimates does not depend on the quality of a model, simply because there is no such model. This is of great importance when these map quality estimates are used as validation statistics, i.e. for checking the validity of the model underlying the map.

In case of non-probability sampling, we need a spatial model of the prediction or classification errors (multivariate distribution of the errors). Several modelling assumptions must be made, such as about stationarity of the mean and of the variance of the errors, and about the covariance of the errors. These assumptions make the estimates prone to discussions. Knotters and Brus (2013) describe an example of how different modelling assumptions about classification errors lead to largely different estimates of the map quality indices. So with non-probability sampling and model-based inference I would not qualify the estimates of the map quality indices and their standard errors a priori as valid estimates. The validity of the modelling assumptions must be carefully checked: is it realistic to assume that the errors are independent, and is it realistic that the mean and or variance of the errors are constant throughout the area? For this reason model-based estimates of map quality indices are less suitable for model (map) validation.

Conclusion

- I define a valid map as a map that is made by a valid model.
- I see map validation as statistical testing of a hypothesis about the validity of the model that is used to construct the map
- The null hypothesis is 'the model is valid'. This null hypothesis cannot be proven. It can be rejected, in which case we conclude that it is very unlikely that the model is valid, or not be rejected, in which case our conclusion is that we do not have evidence that it is not valid.
- The hypothesis can be tested using map quality indices such as the population ME, MSE and SSE as test statistics. The map quality indices then serve as validation statistics.
- When used as validation statistics, the map quality indices can best be estimated by a design-based sampling strategy, involving probability sampling and design based estimation. No model of the prediction or classification errors is used, which guarantees the validity of the estimated map quality indices.

References

Brus, D. J., Kempen, B., and Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. European Journal of Soil Science, 62(3):394–407.

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. Geoderma, 103:3–26.

Knotters, M. and Brus, D. J. (2013). Purposive versus random sampling for map validation: a case study on ecotope maps of floodplains in the Netherlands. Ecohydrology, 6:425– 434.

What’s in a name?

Budiman Minasny

A rose by any other name would smell as sweet… according to a famous author.

I have the benefit of seeing all comments beforehand, but I will still throw in my comments. I agree mostly with Murray that changing a name is not a solution. We still do the same thing, either good or bad validation. Having a new terminology that will be used by a handful of pedometricians would create further confusion.

Hastie et al.’s *Elements of Statistical Learning* (2nd Edition) talk about model selection and model assessment. Model selection: estimating the performance of different models in order to choose the best one. And Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data. They further added that:

“*If we are in a data-rich situation, the best approach for both problems is to randomly divide the dataset into three parts: a training set, a validation set, and a test set. The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model*.”

Most of the time in soil data, we only have a limited number of samples, and thus we cannot divide our data into 3 sets as above. Thus, we do validation or cross-validation. Having a new term “evaluation” does not solve this problem.

Following on Philippe’s clandestine concern, I would like to point out some of the bad habits we like to do on the use of goodness of fit measures:

(1) We like to compare our R^2 or RMSE to other papers to justify our method or results are better. E.g. Our model validation of organic carbon content in such field in Australia has an R^2 of 0.50, which is much higher than the results of Murray (2014) in Scotland who only reported R^2 = 0.30. Clearly you can’t compare that! Unless you are comparing the same field or the same set of data.

(2) We like to justify our validity of goodness of fit by some kind of made-up standard. E.g. Our model has RPD > 2 which means it is accurate and shows a good prediction. Not Necessarily True! There is no basis for such classification, it depends on the data and it is a relative measure.

I would rather that we (including myself) make sure these bad habits are not to be repeated in papers. As Naomi Oreskes in 1994 said “Models can only be evaluated in relative terms”. And the quote I like most from Oreskes is “Verification and validation of numerical models of natural systems is impossible.”