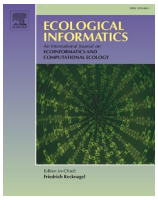
Ecological Informatics 83 (2024) 102820

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Some limitations of the concordance correlation coefficient to characterise model accuracy

Alexandre M.J.-C. Wadoux [a,*], Budiman Minasny [b]

[a] LISAH, Univ. Montpellier, AgroParisTech, INRAE, IRD, L'Institut Agro, Montpellier, France
[b] Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Australia

ARTICLE INFO

Keywords:
Validation
Map quality
Pearson's *r* correlation
Bias
Accuracy
Precision

ABSTRACT

Perusal of the environmental modelling literature reveals that the Lin's concordance correlation coefficient is a popular validation statistic to characterise model or map quality. In this communication, we illustrate with synthetic examples three undesirable statistical properties of this coefficient. We argue that ignorance of these properties have led to a frequent misuse of this coefficient in modelling and mapping studies. The stand-alone use of the concordance correlation coefficient is insufficient because i) it does not inform on the relative contribution of bias and correlation, ii) the values cannot be compared across different datasets or studies and iii) it is prone to the same problems as other linear correlation statistics. The concordance coefficient was, in fact, thought initially for evaluating reproducibility studies over repeated trials of the same variable, not for characterising model accuracy. For the validation of models and maps, we recommend calculating statistics that, combined with the concordance correlation coefficient, represent various aspects of the model or map quality, which can be visualised together in a single figure with a Taylor or solar diagram.

1. Introduction

The quality of predictions in environmental modelling and mapping studies is usually determined through the pairwise comparison of measured/observed and predicted values, from which summary validation statistics describing the overall correspondence are calculated. Common validation statistics are the mean error, mean absolute error, the root mean square error, the R^2 and the modelling efficiency coefficient. Perusal of the literature reveals that the concordance correlation coefficient (ρ_c, Lin, 1989) is another popular statistics to evaluate the overall model or map quality. Zhao et al. (2022), for example, used the ρ_c to compare predictions of soil clay at field scale obtained from multiple sensors, whereas in Caubet et al. (2019) the ρ_c was used to compare national, continental and global maps of soil texture in France. It is also a popular validation statistics in environmental simulation and sensitivity analysis studies (e.g. Branco et al., 2006; Lim et al., 2018). Chapagain et al. (2023), for example, used the the ρ_c to compare model outputs generated by different crop simulation model structures. While estimating the quality of prediction using adequate validation statistics has been the purpose of many studies in ecological modelling and in the broader applied statistics literature (see, for example, Janssen and Heuberger (1995) and Power (1993)), little has been described of the statistical properties of the ρ_c for validating the prediction of environmental models and for assessing the quality of maps.

Lin (1989) proposed that the ρ_c accounts for both precision and bias when evaluating the agreement from trial to trial in validation and measurement reproducibility studies. The ρ_c evaluates the degree to which pairs of observations fall on the 45-degree line through the origin in a scatterplot, addressing the limitation of the linear correlation coefficient. In environmental modelling studies, difficulties arise because the ρ_c is frequently assumed to provide a single measure of model or map quality. For example, many of the soil modelling studies using the ρ_c have taken thresholds values provided by Viscarra Rossel and Hicks (2015) to assess the reliability of the predictions (i.e. > 0.90 is an excellent agreement, $0.80 < \rho_c < 0.90$ is a substantial agreement, $0.65 < \rho_c < 0.8$ is a moderate agreement, and values below 0.65 represent poor agreement), whereas McBride (2005) in the broader environmental sciences literature suggested different thresholds (i.e. > 0.99 is almost perfect agreement, $0.95 < \rho_c < 0.99$ is a substantial agreement, $0.90 < \rho_c < 0.95$ is a moderate agreement, and values below 0.9 represent poor agreement), and Altman (1990) gave another scale (i.e. $0.81 < \rho_c < 1$ is a very good agreement, $0.61 < \rho_c < 0.80$ is a good agreement, $0.41 <$

* Corresponding author at: Laboratory of Soil-Agrosystem-Hydrosystem Interaction (LISAH), 2 Place Pierre Viala, 34090 Montpellier, France.
E-mail address: alexandre.wadoux@inrae.fr (A.M.J.-C. Wadoux).

https://doi.org/10.1016/j.ecoinf.2024.102820
Received 21 February 2024; Received in revised form 4 September 2024; Accepted 7 September 2024
Available online 11 September 2024

$\rho_c < 0.60$ is a moderate agreement, $0.21 < \rho_c < 0.40$ is a fair agreement and values below 0.2 represent poor agreement). These thresholds, however, have provoked a general confusion and several misapplications of the ρ_c because the values vary dramatically in response to the variability in the dataset. While the limitation of the ρ_c have been discussed in the literature (e.g. Atkinson and Nevill, 1997), we underline that in the absence of the critical evaluation, the Lin's ρ_c could be misused in environmental sciences research. In this note, we highlight three limitations and undesirable statistical properties of the ρ_c, which have been described in the literature and acknowledged by later work of Lin, such as Lin et al. (2002), but insufficiently recognised by its users. We illustrate our point with elementary examples using synthetic datasets.

Consider two vectors of values, hereafter referred to as measured and predicted values and denoted $\mathbf{z}$ and $\widehat{\mathbf{z}}$, respectively, with mean $\overline{z}$ and $\overline{\widehat{z}}$ and standard deviation σ_z and $\sigma_{\widehat{z}}$. The concordance correlation coefficient (ρ_c, Krippendorff, 1970; Lin, 1989) quantifies the agreement of two sets of values by scaling in the range -1 to 1 the expected value of the squared perpendicular distance from the diagonal $\mathbf{z} = \widehat{\mathbf{z}}$ (i.e. the 45^0 line through the origin). Lin's ρ_c is given by:

$$\rho_c = \frac{2r\sigma_{\widehat{z}}\sigma_z}{\sigma_z^2 + \sigma_{\widehat{z}}^2 + (\overline{z} - \overline{\widehat{z}})^2}, \tag{1}$$

where $r\sigma_{\widehat{z}}\sigma_z$ is the covariance between observed and predicted values and r is a linear correlation coefficient. Analogous to the Pearson's r correlation coefficient, a value of 1 indicates perfect agreement and -1 perfect disagreement. Further aspects of the relationship between r and ρ_c is shown in Lin (1989) by reducing Eq. (1) to rC_b where C_b is a bias correction factor defined as:

$$C_b = \left(\frac{\sigma^* + \frac{1}{\sigma^*} + u^2}{2} \right)^{-1}, \tag{2}$$

with

$$\sigma^* = \frac{\sigma_{\widehat{z}}}{\sigma_z}; \quad u = \frac{\overline{\widehat{z}} - \overline{z}}{\sqrt{\sigma_{\widehat{z}}\sigma_z}}. \tag{3}$$

From Eqs. (2) and (3) it follows that $C_b = 1$ when all the points are on the diagonal $\mathbf{z} = \widehat{\mathbf{z}}$. The more the points deviate from the diagonal line, the closer C_b is to 0, thus $0 \leqslant C_b \leqslant 1$. It is further considered that σ^* is a measure of multiplicative shift, and u a measure of additive shift relative to a multiplicative shift. The reduction of the concordance correlation coefficient to rC_b makes clear that ρ_c has always the sign of r, and that $\rho_c = 0$ if $r = 0$. The ρ_c is thus a measure of both precision and accuracy, which are characterised by the r and C_b, respectively. We hypothesise that this is likely the reason why so many researchers adopted the concordance correlation coefficient as a single measure of model or map quality and for making comparisons among studies on different populations.

First limitation: The ρ_c does not inform on the individual contribution of correlation and bias. Consider Fig. 1 with four maps, one of which is a reference and three are modifications of the reference considered as prediction. The reference map was obtained through simulation on a regular grid of 100 × 100 cells. The simulation had a linear trend superimposed on a Gaussian random field. The trend had an intercept of 5, a slope of 0.1 for the x-axis and a slope of 0.05 for the y-axis. The Gaussian field had a mean of zero and an exponential covariance function with a range of 10 and a sill of 10. The three prediction maps have a very different spatial pattern. The maps and scatterplots between the reference and prediction maps show that prediction 1 has a pattern smoother than the reference with a strong underestimation of large values. Prediction 2 has a systematic overestimation of the reference values, which becomes more important for larger values. Prediction 3 has an underestimation of small values and a strong overestimation of large values. In each of these cases, however, the ρ_c shows the exact same level of agreement with the reference map despite substantial differences in the individual contributions of correlation and bias (Prediction 1: $\rho_c = 0.6, r = 0.69, C_b = 0.87$; Prediction 2: $\rho_c = 0.6, r = 1, C_b = 0.6$; Prediction 3: $\rho_c = 0.6, r = 0.66, C_b = 0.91$). From the ρ_c and

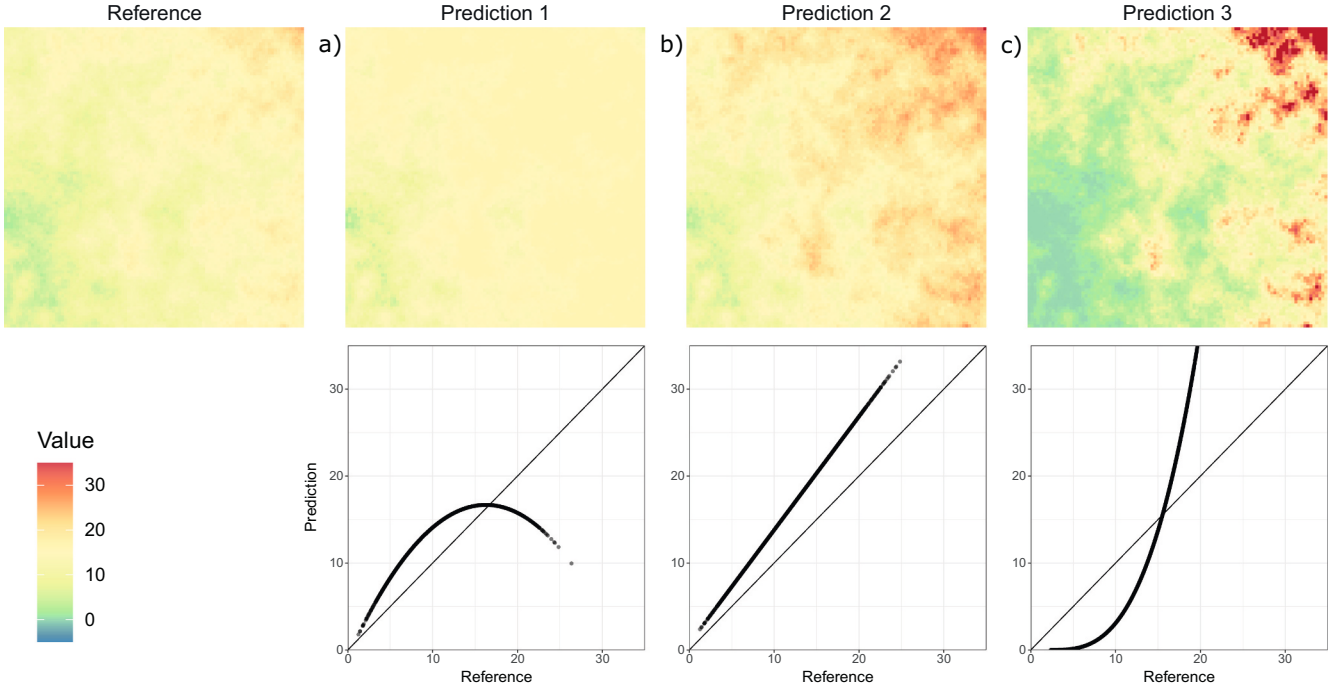


Fig. 1. A reference map with three modifications of this map taken as prediction. Predictions 1 has predicted values $\widehat{z}_1 = -0.82 + 2.15z - 0.07z^2$, where z is the value from the reference map. Prediction 2 is obtained with $\widehat{z}_2 = 0.74 + 1.31z$ and prediction 3 with $\widehat{z}_3 = 0.007(z - 2.248)^3$. The three prediction maps have the exact same level of agreement (i.e. $\rho_c = 0.6$) with the reference map.

without further information one cannot know if the disagreement between the reference and predicted maps is due to bias or lack of precision.

Second limitation: The values can not be compared across studies since the ρ_c values are sensitive to the variance of the measured data. Consider Table 1 which shows two datasets, each with a set of 20 measured and associated predicted values. Measured and predicted values are simulated from a normal distribution with a mean of 20 and standard deviation of 2 in the first dataset. For the second dataset, measured values were simulated from a normal distribution with the same mean but standard deviation of 6, and the prediction are obtained by adding up the simulated measured values to the residuals of the first dataset. The first dataset has a standard deviation of 1.65 and 1.79 for the measured and predicted values, respectively, and show little agreement with a ρ_c value of 0.21. In the second dataset, the measured values are less homogeneous with a standard deviation value of 6 but with the same residuals between measured and predicted values as the first dataset. This time, however, the ρ_c is 0.93, showing a nearly perfect agreement. This is only due to the difference in standard deviation between the first and second datasets.

Third limitation: The ρ_c is prone to the same problems as other linear correlation statistics. Fig. 2 illustrates three cases of predictions which deviate from the line of equality and contain errors. Fig. 2a shows predictions with a multiplicative shift, Fig. 2b has three predictions with a $(1-\beta_1)z-\beta_0$ deviation from the line of equality, for different values of β_0 and β_1, whereas Fig. 2c shows three predictions with negative and positive systematic shifts from the measurements. All predictions have a perfect linear correlation with the measurements with a corresponding value of 1. Note that since in all cases the Pearson's r correlation coefficient is 1 the ρ_c takes the value of the bias correction factor C_b. It was shown in the literature (e.g. Willmott, 1984) that the correlation coefficient is insensitive to additive and proportional difference between measured and predicted values and is not a good measure of prediction accuracy because it is unrelated to the size of the error (Li, 2017; Willmott, 1982). Fig. 1 showed some well-known examples of predictions having perfect correlation despite having large difference with the measurements.

The results show that while the ρ_c is a single index accounting for both precision and bias, different predictions may lead to the same value of ρ_c and one cannot distinguish the individual contribution of bias and correlation. This has important implication because without further information (e.g. using the mean error) it is not possible to discern whether a low value of ρ_c is due to a systematic deviation or random error between measured and predicted value (Atkinson and Nevill, 1997).

The ρ_c is further subject to the same problems as those inherent in the linear correlation statistics: its value depends on the variability in the data. For this reason, it is easy to obtain a high value of ρ_c with heterogeneous datasets. This means that the ρ_c should not be used as a single validation statistic informing on the relative accuracy of a map or model over another if the measured data used as reference are not the same. These limitations also highlight that existing threshold values interpreting the ρ_c as "excellent" or "poor". which are not based on any statistical or utilitarian basis, are misleading and can lead to wrong conclusion on the quality of predictions is modelling studies.

While we stress here that the stand-alone use of the ρ_c is not a reliable way of assessing model or map accuracy beyond the comparison of maps or models against a baseline on the same target population, the question that arises is which validation statistics are efficient to characterise the output the a prediction model. There has been several attempts (e.g. King and Chinchilli, 2001; Leal et al., 2019) to correct for the limitations of the ρ_c. For example, Vallejos et al. (2020) introduced a new coefficient to assess the concordance between spatial variables, under different correlation structures and variances, thus addressing the limitations explained in the first experiment. While these are worthwhile efforts, no single validation statistic can represent all aspects of the model or map quality. We suggest evaluating the quality of model or maps using complementary indices; an index that is sensitive to the deviation from the 1:1 line, such as the modelling efficiency coefficient (Janssen and Heuberger, 1995), as well as indices of bias (e.g. the mean error) and magnitude of errors (e.g. the mean absolute or squared error) (Willmott, 1984). Several publications advocate for a similar combination of statistics. Moriasi et al. (2007), for example, recommended the Nash-Sutcliffe efficiency, the percent bias and the ratio of the root mean square error to the standard deviation of measured data. Another example is Caubet et al. (2019), where multiple validation statistics are

Table 1
Two datasets with similar mean but different variability. The two datasets have the same residuals between the measured and predicted values. The first dataset has a prediction with a $\rho_c = 0.21$ ($C_b = 0.86$ and $r = 0.25$) while the second has a $\rho_c = 0.94$ ($C_b = 0.99$ and $r = 0.95$). Simulated example adapted from Atkinson and Nevill (1997).

First dataset			Second dataset		
Measurements	Prediction	Residuals	Measurements	Prediction	Residuals
21	21	0	37	37	0
19	18	1	19	18	1
22	19	3	19	16	3
21	20	1	21	20	1
22	22	0	23	23	0
19	24	−5	10	15	−5
18	20	−2	10	12	−2
20	21	−1	24	25	−1
16	19	−3	24	27	−3
19	23	−4	24	28	−4
20	22	−2	15	17	−2
21	22	−1	17	18	−1
21	21	0	18	18	0
20	20	0	12	12	0
21	23	−2	19	21	−2
21	18	3	25	22	3
21	24	−3	11	14	−3
22	22	0	25	25	0
19	20	−1	20	21	−1
17	20	−3	27	30	−3
Mean					
20	20.95	−0.95	20	20.95	−0.95
Standard deviation					
1.65	1.79	2.11	6.62	6.39	2.11

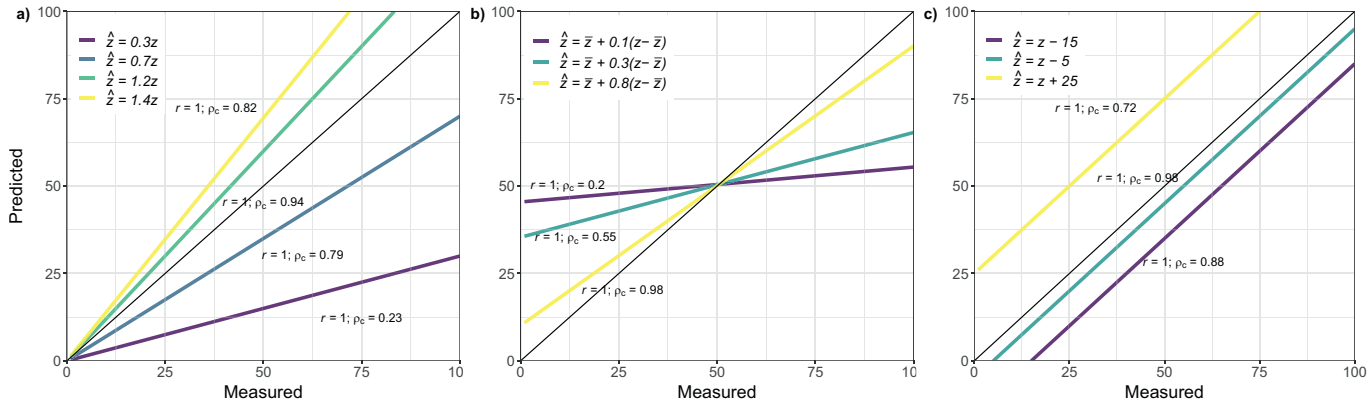


Fig. 2. Plots with between measured and predicted values for three cases, for predictions a) with a multiplicative shift from the measurements, b) obtained through $\hat{z} = \bar{z} + \beta(z - \bar{z})$, and c) with a systematic over- or under-estimation. The colours indicate different coefficient values.

used in combination with the ρ_c. We suggest that, in addition to using multiple validation statistics, one can benefit from the statistical relationship between indices and plot them together in a single figure. Taylor and solar diagrams (Wadoux et al., 2022) can be used for this purpose. The solar diagram enables the direct visualization of the ρ_c together with the standard deviation of the error, the mean error and the linear correlation. Such plots can be complemented by a residual or prediction plot.

CRediT authorship contribution statement

Alexandre M.J.-C. Wadoux: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Budiman Minasny:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization.

Data availability

The experiments conducted in this study along with the simulated data are publicly available at: https://github.com/AlexandreWadoux/LCCC_paper_experiments.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101059012. For the purpose of Open Access, a CC-BY public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission.

References

Altman, D.G., 1990. Practical Statistics for Medical Research. CRC Press, Boca Raton.
Atkinson, G., Nevill, A., 1997. Comment on the use of concordance correlation to assess the agreement between two variables. Biometrics 53, 775–777.
Branco, M., Jactel, H., Franco, J.C., Mendel, Z., 2006. Modelling response of insect trap captures to pheromone dose. Ecol. Model. 197, 247–257.
Caubet, M., Dobarco, M.R., Arrouays, D., Minasny, B., Saby, N.P.A., 2019. Merging country, continental and global predictions of soil texture: lessons from ensemble modelling in France. Geoderma 337, 99–110.
Chapagain, R., Huth, N., Remenyi, T.A., Mohammed, C.L., Ojeda, J.J., 2023. Assessing the effect of using different apsim model configurations on model outputs. Ecol. Model. 483, 110451.
Janssen, P.H.M., Heuberger, P.S.C., 1995. Calibration of process-oriented models. Ecol. Model. 83, 55–66.
King, T.S., Chinchilli, V.M., 2001. Robust estimators of the concordance correlation coefficient. J. Biopharm. Stat. 11, 83–105.
Krippendorff, K., 1970. Bivariate agreement coefficients for reliability of data. Sociol. Methodol. 2, 139–150.
Leal, C., Galea, M., Osorio, F., 2019. Assessment of local influence for the analysis of agreement. Biom. J. 61, 955–972.
Li, J., 2017. Assessing the accuracy of predictive models for numerical data: not r nor r2, why not? Then what? PLoS One 12, e0183250.
Lim, R.B.H., Liew, J.H., Kwik, J.T.B., Yeo, D.C.J., 2018. Predicting food web responses to biomanipulation using Bayesian belief network: assessment of accuracy and applicability using in-situ exclosure experiments. Ecol. Model. 384, 308–315.
Lin, L.I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268.
Lin, L., Hedayat, A.S., Sinha, B., Yang, M., 2002. Statistical methods in assessing agreement: models, issues, and tools. J. Am. Stat. Assoc. 97, 257–270.
McBride, G.B., 2005. A Proposal for Strength-of-Agreement Criteria for Lin's Concordance Correlation Coefficient. Technical Report National Institute of Water & Atmospheric Research Hamilton, New Zealand. NIWA Client Report: HAM2005–062.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50, 885–900.
Power, M., 1993. The predictive validation of ecological and environmental models. Ecol. Model. 68, 33–50.
Vallejos, R., Pérez, J., Ellison, A.M., Richardson, A.D., 2020. A spatial concordance correlation coefficient with an application to image analysis. Spat. Stat. 40, 100405.
Viscarra Rossel, R.A., Hicks, W.S., 2015. Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. Eur. J. Soil Sci. 66, 438–450.
Wadoux, A.M.J.-C., Walvoort, D.J.J., Brus, D.J., 2022. An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. Geoderma 405, 115332.
Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63, 1309–1313.
Willmott, C.J., 1984. On the evaluation of model performance in physical geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial Statistics and Models. Springer, Dordrecht, NL, pp. 443–460.
Zhao, X., Wang, J., Zhao, D., Triantafilis, J., 2022. Soil organic carbon prediction by multi-digital data fusion for nitrogen management in a sugarcane field. Nutr. Cycl. Agroecosyst. 127, 1–18.