General

The authors introduce "Monte Carlo conformal prediction" (MC-CP), a combination of Monte Carlo dropout and conformal prediction, for the uncertainty quantification of soil spectroscopy predictions with deep learning models. They demonstrate its merits compared to pure Monte Carlo dropout or conformal prediction approaches for in-domain predictions. Furthermore, the paper is logically structured and easy to follow. However, the introduction describes several concepts incompletely. Moreover, I do not share their conclusions regarding the "out-of-domain" predictions, as they claim that MC-CP can address uncertainty for out-of-domain data, even though all they show is that the prediction intervals are slightly wider for out-of-domain data, without providing information on the (presumably poor) coverage. These key issues need to be addressed prior to publication. See below specific comments:

 # Comment 1; L. 7

The abstract should mention that Monte Carlo dropout is a method for neural networks (i.e., deep learning). Currently, it mentions “machine learning“ in L. 7, which could appear as if the method is model-agnostic for (any) machine learning model. Either, machine learning could be replaced with deep learning, or it could be somewhere else explicitly mentioned that it is a method for deep learning.

 # Comment 2; L. 12

The abstract should mention that Monte Carlo dropout is a method for neural networks (i.e., deep learning). Currently, it mentions “machine learning“ in L. 7, which could appear as if the method is model-agnostic for (any) machine learning model. Either machine learning could be replaced with deep learning, or it could be explicitly mentioned somewhere else that it is a method for deep learning.

Literature

Bethell, D., Gerasimou, S., & Calinescu, R. (2024). Robust uncertainty quantification using conformalised Monte Carlo prediction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 19, pp. 20939-20948).

Romano, Y., Patterson, E., & Candes, E. (2019). Conformalized quantile regression. Advances in neural information processing systems, 32.

 # Comment 3; L. 20 -21

I do not share the opinion that MC-CP “effectively address[ed] the higher uncertainty in out-of-domain samples” given the results presented in this paper but my discussion on that can be found in Comment 19.

 # Comment 4; L. 23- 26

I am convinced of the merits of MC-CP for deep learning in soil spectroscopy but the wording is exaggerated here. Neither “breakthrough” nor “revolutionizing” are appropriate terms here because the authors did not invent this method but demonstrated its advantage compared to their vanilla version for “in-domain data”.

 # Comment 5; L. 28-29

In the recent developments of soil science, machine learning has been widely used, such as soil spectroscopy, proximal sensing, carbon stock modelling, and digital soil mapping (Padarian et al., 2020; Minasny et al., 2024). Wording around "such as" sounds slightly off, and I propose adding "in applications": "[..] widely used in applications such as soil spectroscopy, proximal sensing […]"

 # Comment 6; L. 50 – 67

The concept of aleatoric and epistemic uncertainty is quite vague and may even be incorrectly applied here because of the sentence in L. 54: "Epistemic uncertainty is the main topic in this study."

In the following, I use the definitions of Valdenegro-Toro & Mori (2022): "*There are two kinds of uncertainty […]: aleatoric or data uncertainty, and epistemic or model uncertainty. These uncertainties are usually combined and predicted as a single value, called predictive uncertainty […].*" Hence, the interest of the study is to find an uncertainty method (e.g. MC-CP) which succeeds in quantifying the combined predictive uncertainty.

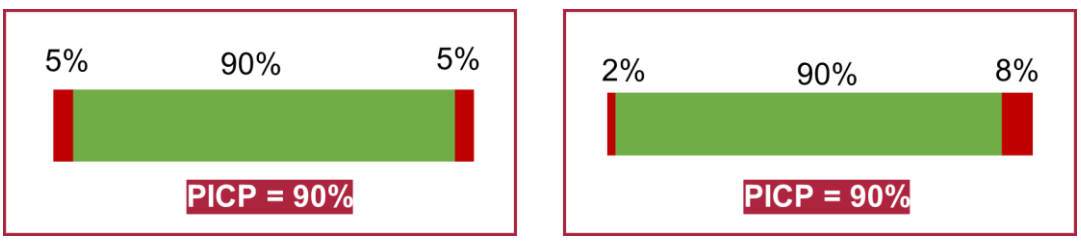
Of course, epistemic uncertainty becomes especially relevant for the "out-of-domain data" predictions, as here epistemic uncertainty is very high. Hence, it is relevant that the uncertainty quantification model can also account for the epistemic uncertainty that is associated with the domain shift. Maybe this is what the authors intended to refer to but it is a far stretch to what is written and nowhere explicitly mentioned. Currently, it appears as if the authors confuse the predictive uncertainty with the epistemic uncertainty.

Literature

Valdenegro-Toro, M., & Mori, D. S. (2022, June). A deeper look into aleatoric and epistemic uncertainty disentanglement. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1508-1516). IEEE.

 # Comment 7; L. 57 – 58:

Strictly speaking, this is inaccurate. An ideal uncertainty quantification method should have ideal coverage of the quantiles. Imagine a 90% prediction interval, which consists of a 5% and 95% quantile. So, an ideal uncertainty quantification method should have a 5% and 95% coverage of the quantiles. If the 5% quantile is covered by 2% of the test samples, and the 95% quantile by 92% of the test samples, the PICP would be still 90%, even though the uncertainty was wrongly quantified. See for example the short paper on PICP from Pinson & Tatsu (2014) or figure below.



Literature

Pinson, P., & Tastu, J. (2014). Discussion of “Prediction intervals for short-term wind farm generation forecasts” and “Combined nonparametric prediction intervals for wind power generation”. IEEE Transactions on Sustainable Energy, 5(3), 1019-1020.

 # Comment 8; L. 61 – 66

It is hard to understand why the authors mention only bootstrapping in this context. It is correct, that bootstrapping is used occasionally in soil science to quantify uncertainties. However, this is a common methodological error because bootstrapping only creates “confidence intervals” not “prediction intervals” i.e., pure bootstrapping was never intended to be used to quantify predictive uncertainties. On the other hand, the authors leave out the most commonly used method in soil: quantile regression (e.g., quantile regression forest, XGBoost with quantile loss function etc.) or even conformalized quantile regression. Quantile regression is not (yet?) well implemented for deep learning, which is why MC-CP becomes relevant but the introduction feels incomplete in the context of soil. More so, because conformalized quantile regression was recently introduced in soil by Kakhani et al. (2024), which is in its logic very related to MC-CP.

Literature

Kakhani, N., Alamdar, S., Kebonye, N. M., Amani, M., & Scholten, T. (2024). Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction. Remote Sensing, 16(3), 438.

 # Comment 9; L. 76 – 86

The concept of aleatoric and epistemic uncertainty may be applied to this section.

 # Comment 10; L. 93 – L. 95

One may argue that quantile regression could do so too, which is why it needs to be discussed somewhere earlier.

 # Comment 11; L. 96

More appropriate would be to replace “we applied a strategy to increase the PICP of MC dropout” with “we applied a strategy to improve the PICP coverage of MC dropout”, unless MC generally leads to too narrow prediction intervals and not suboptimal coverage.

 # Comment 12; L. 118 – 120:

During my first read, I was wondering how the testing and validation was done. It is mentioned in L. 203 but for better readability it could be already defined here. I agree with it either way.

 # Comment 13; L. 123

It may be considered to use “it has been proven” instead of “it has been proved”.

 # Comment 14; L. 134:

Sounds slightly off. One suggestion to highlight that the predictive distribution is inferred from 100 trained CNN models with dropout layers: “In practice, a CNN model with dropout layers was trained 100 times to generate a predictive distribution”.

 # Comment 15; L. 135-136:

The 90 % prediction interval is not the difference between the 5^{th} and 95^{th} quantile, but the prediction interval itself is the interval defined by those two quantiles. The difference is the “(mean) prediction interval width”. In Eq. 1 it is shown correctly, meaning it is just an issue of terminology.

 # Comment 16; Eq. 1/ L.137:

C is the 90% prediction interval, which is relatively logical given the previous sentence. Nonetheless, it could be defined in the text. Also, if it is called C_{90}, it would follow the scheme of $\hat{q}_5$ and $\hat{q}_{95}$.

 # Comment 17; L. 222:

Wording could be considered: it is correct that it is expectable that the model fails to do proper out-of-domain predictions, but the performance is not only “poor” but completely unusable since the R^2 = -6.64. The word “poor” indicates to me a model with an R^2 around or slightly above 0.

 # Comment 18; L. 250/Fig. 3:

It seems a bit incoherent that the plot covers the 0 – 90% range instead of 0 – 100% range.

 # Comment 19; Out-of-domain results and discussion:

I do not fully agree with the discussion on the out-of-domain results for the following reasons:

First of all, it is not clear why the authors do not show the PICP for the out-of-domain predictions, presumably because it has very poor coverage, given that the MPIW is only marginally larger even though the model is extremely bad for out-of-domain predictions. A much larger uncertainty (i.e., MPIW) should be expected here.

Instead, the authors focus on the fact that the MPIW is slightly larger for out-of-domain predictions with MC-CP and MC compared to in-domain predictions. It correctly shows that the model is somewhat aware that it is less certain for out-of-domain data. The authors see this as a reason to conclude that MC-CP is able to address the uncertainty of out-of-domain samples (L. 278-281). However, the results do not really support this claim because the MPIW alone is uninformative without information on the coverage. I highly assume that the MPIWs are still not wide enough. For extensive conclusions, the authors should include the coverage for the out-of-domain predictions. Hence, the coverage of the out-of-domain data needs to be included as well!

A second proceeding problem may occur if the PICP is used for evaluating out-of-domain samples, and it is strongly associated with the previous comment 7.

It can be expected that the observed values will be much more likely to be above the 95% quantile (as shown in the right example of Fig. 2!) and less often below the 5% quantile because the observed out-of-domain values have higher clay values than the model is trained on. Hence, evaluating the quantiles would make much more sense than using the PICP. This is the more standard practice and has been addressed in the context of soil too (Schmidinger & Heuvelink, 2023).

Literature

Schmidinger, J., & Heuvelink, G. B. (2023). Validation of uncertainty predictions in digital soil mapping. Geoderma, 437, 116585.

 # Comment 20; L. 255:

MPIW instead of PIW.

 # Comment 21; L. 344:

"The authors benefited from the shared code of Daniel Bethell. The manuscript would become much more impactful for the soil community if the authors shared their code as well. This would increase the usability of the MC-CP method. Given that some KSSL data has been published in OSSL, it would also be easy to reproduce the study."