Machine learning to generate soil information

José Sergei Padarian Campusano

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

2019

Faculty of Science

School of Life and Environmental Sciences

The University of Sydney

New South Wales

Australia

Certificate of Originality

This thesis is submitted to the University of Sydney in fulfilment of the requirements for the degree of Doctor of Philosophy.

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

Signature: José Padarian

Date: April 2, 2020

Thesis summary

This thesis is concerned with the novel use of machine learning (ML) methods in soil science research. Since the first applications of ML methods in the 80s, ML adoption in soil science has increased considerably, especially in pedometrics (the use of quantitative methods to study the variation of soils). In parallel, the size of the soil datasets has also increased thanks to projects of global impact that aim to rescue legacy data or new large extent surveys to collect new information. While we have big datasets and global projects, currently, modelling is mostly based on "traditional" ML approaches which do not take full advantage of these large data compilations. In some cases, datasets from different parties are compiled into a single source and used by a single organisation to create models. This process is severely limited by privacy concerns and, currently, no solution has been implemented to facilitate the process. In the context of digital soil mapping (DSM), while there are calls for "bottom up" approaches (e.g. GlobalSoilMap), they have not been generally applied, and they have technical challenges associated such as harmonisation. If we also consider the performance differences derived from the generality of global models versus the specificity of local models, there is still a debate on which approach is better. Either in global or local DSM, most applications are static. Even with the large soil datasets available to date, there is not enough soil data to perform a fully-empirical, space-time modelling. The application of ML in soil sciences usually prioritise numerical or categorical data over qualitative descriptions. Considering the resources that have been invested in collecting large amounts of descriptive information, neglecting this data seems wasteful, yet methods that take advantage of this type of information have rarely been applied in soil sciences.

Considering these knowledge gaps, this thesis aims to introduce advanced machine learning algorithms and training techniques, specifically deep neural networks, for modelling large datasets at a global scale and provide new soil information. The following paragraphs introduce some of the techniques utilised in this thesis.

Chapter 1 explores the use of ML in soil sciences by means of a systematic review aided by machine learning tools. The main outcome is the result of a topic modelling analysis, which is a probabilistic ML method that aims to discover and annotate large archives of documents with thematic information, assigning each document a

probability of belonging to a specific topic. This method allows processing of a large number of articles, which can help to reduce part of the bias introduced by only selecting a manageable subset of documents, or by manually assigning documents to topics.

As shown in Chapter 1, DSM is a widely adopted discipline, which has embraced the use of ML techniques. However, there are some aspects of DSM which have been recognised by the scientific community as future improvements of the framework. Chapter 2 extends the DSM approach by proposing two improvements. The first one is related with the spatial context of a soil observation. Most DSM studies are based on point-support (profiles with point coordinates), from where the scorpan variables (environmental information used as proxies for soil forming factors) are extracted (pixel intersected by the point). This is equivalent to a soil scientist only observing what it is inside a soil pit. In reality, when a soil scientists analyses a soil pit also consider its surroundings. In this chapter we propose the use of an image instead of a single pixel — a window around the point observation. Since most models used in DSM are not able to process this information, we introduce the use of deep convolutional neural networks (CNN) in soil sciences. With the proposed approach we were able to reduce the error by 30% compared with conventional techniques. The second limitation is related to the prediction in depth. As a consequence of mostly using covariates that represent surface conditions, DSM usually shows a decrease in the variance explained by the model as the prediction depth increases. But, if we are capable of making good predictions for the top layers, can we use that information to predict in depth? With the proposed CNN, we are capable of simultaneously optimising the predictions of surface and sub-surface layers, proposing an effective method to exploit the relation between multiple depths or soil properties.

Due to intrinsic cost of field surveys, obtaining soil information is usually a limitation, especially for global scale projects. In this thesis, this challenge is approached from two perspectives: find methods to promote collaboration between researchers and improve methods to obtain soil information, faster and cheaper. In order to foster collaboration, Chapter 3 proposes a ML method to generate a single model where the data remains with each party and there is no need to integrate it to a single source. This approach can help dissipate concerns of data privacy and confidentiality that are still an obstacle that impedes progress in collaborative global research, despite the well-known benefits of collaboration. This approach is

demonstrated by building a global soil organic carbon model based on databases of field observations held by 65 different countries. The model is trained by visiting each country, one at a time. Only knowledge and parameters of the model are transferred between countries. The results show that it is possible that the proposed approach yields a similar prediction accuracy compared with a model that is trained with all the data.

One of the methods that is routinely used to generate soil information thanks to its speed and cost-effectiveness is NIR spectroscopy. In Chapter 4, we propose the novel use of CNNs to achieve two purposes: a) avoid spectral pre-processing and b) predict multiple properties simultaneously. The multi-task model was trained using a large continental soil database (LUCAS) and compared to two models traditionally used to predict soil properties from spectral data, namely PLS regression and Cubist. The multi-task CNN model outperformed the traditional models when predicting six soil properties, notably by 61.6 and 55.4% for organic carbon and total nitrogen. This dramatic improvement is rarely found in spectroscopy studies and it was possible thanks to the simultaneous (with one model) prediction of six soil properties from a single spectrum, in a truly synergistic approach.

One of the conclusions derived from the results of Chapter 4 was that the method is only effective in the presence of a large database. Deep learning methods (neural networks with multiple layers) are well-known for being a "data-hungry" approach and, in practice, collating a large soil spectral database is not a simple task. In addition, general models like the one presented in Chapter 4 usually perform poorly when applied at the local scales. In order to take full advantage of large, publicly available datasets, which are composed of a variety of soil types and conditions, a method that extracts part of the general "experience" contained in that data to make predictions at the local scale is desirable. Chapter 5 presents a solution to connect multiple modelling scales by evaluating the effectiveness of transfer learning to "localise" a general soil spectral model. The process consists of training a general model, which is then dissected, extracting some layers of the neural network to build a local model. This local model is then fine-tuned using a smaller, local dataset. We tested the approach using the LUCAS dataset to generate 21 country-specific models to predict organic carbon, cation exchange capacity, clay content and pH, simultaneously. The resulting "localised" models outperformed the local and general model (trained only with local or complete

datasets, respectively) in 91% of the cases demonstrating that there was a positive transfer of knowledge between scales.

In Chapter 1, we explored part of the capabilities of natural language processing. In Chapter 6 we introduce the use of word embedding for geosciences. Word embeddings are a numerical representation of words which can be then used in traditional numerical analysis, taking advantage of the large volume of descriptive data available which is usually disregarded. A language model was generated taking into account the co-occurrences of words within a large corpus of 280,764 full-text scientific articles related to geosciences. The resulting model generates a multi-dimensional vector space where angles and distances have a linguistic interpretation. In a practical example, we emulated part of an analysis related to soil numerical taxonomy. The results obtained by using our domain-specific embeddings to represent soil profiles descriptions were equivalent to the results derived from numerical laboratory data.

Despite the large, global soil datasets available to date, there is still a lack of information to apply ML methods to solve specific modelling challenges. In Chapter 7, we present an example of a global space-time assessment of soil organic carbon where, even if we use the largest soil dataset compiled to date, it is not possible to generate a fully-empirical, data-driven model. We propose incorporating a) a machine learning model to link environmental covariates with soil carbon contents, b) a mechanistic component that emulates carbon dynamics dependent on precipitation and temperature, and c) a landcover-tracking component that varies the outputs depending on the landcover history. We estimated a global carbon stock (with bulk density inferred using a pedotransfer function) in the top 30 cm of around 793.16 Pg with annual losses due to landcover change of 1.93 Pg SOC yr^{-1} between 2001 and 2016. The biggest losses are concentrated in the tropic and sub-tropical regions, accounting for almost 50% of the total loss (0.9 Pg yr^{-1}). Our proposed modelling framework is flexible, allowing it to be updated as more or better data becomes available.

The research presented here has been successful at applying the latest advances in ML to improve upon some of the current approaches for soil modelling large datasets at large scales and provides new soil information. It has also created opportunities to utilise information, such as descriptive data, that has been generally disregarded. ML methods have been embraced by the soil community and their adoption is increasing. In the particular case of neural networks, their flexibility in terms of structure and training

makes them a good candidate to improve on current soil modelling approaches.

Chapters of this thesis that have been submitted and/or published in scientific journals

Chapter 1

Padarian, J., Minasny, B., and McBratney, A. B. 2020. Machine learning and soil sciences: a review aided by machine learning tools, Soil, 6, 35–52.

Chapter 2

Padarian, J., Minasny, B. and McBratney, A.B., 2019. Using deep learning for digital soil mapping. Soil, 5(1), pp.79-89.

Chapter 3

Padarian, J., Minasny, B. and McBratney, A.B., 2019. Online machine learning for collaborative biophysical modelling. Environmental Modelling & Software, 122, p.104548..

Chapter 4

Padarian, J., Minasny, B. and McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. Geoderma Regional, 16, p.e00198.

Chapter 5

Padarian, J., Minasny, B. and McBratney, A.B., 2019. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma, 340, pp.279-288.

Chapter 6

Padarian, J. and Fuentes, I., 2019. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. Soil, 5(1), pp.177-187.

Contents

List of Figures

List of Tables

General introduction

Machine learning (ML) is a sub-discipline of artificial intelligence (AI) which studies algorithms that perform tasks without having programmatical, pre-defined rules. ML algorithms rely on learning patterns that are inferred from training data, usually in an iterative process which can be assimilated as experiential. According to some authors, ML lays in the intersection between statistics and computer sciences (Jordan and Mitchell, 2015). In soil sciences, statistical, AI and ML methods have been widely applied. The use of statistical methods to predict soil properties dates back to early 20th century (Briggs and McLane, 1907; Briggs and Shantz, 1912). During the 80s, expert systems (an AI system that emulates the human decision-making process based on pre-defined rules) became popular (Dale *et al.*, 1989), but given the difficulty to pre-generate most of the rules, usually by domain experts, they were displaced by methods that automatically generated rules based on data, such as early machine learning algorithms (e.g. CART (Breiman *et al.*, 1984)).

Since the first applications of ML methods during the 80s, ML adoption in soil science has increased considerably, especially in sub-disciplines such as pedometrics. Pedometrics, defined as the use of quantitative methods to study the variation of soils (Burrough *et al.*, 1994), includes two disciplines that have embraced ML methods: digital soil mapping (DSM) and soil spectroscopy. DSM has been applied all over the world and it is the backbone of many highly publicised initiatives such as GlobalSoilMap (Arrouays *et al.*, 2014), 4per1000 (Minasny *et al.*, 2017) and FAO's Global Soil Organic Carbon Map. Soil spectroscopy has proved to be an inexpensive alternative to conventional laboratory analysis when characterising soil properties. Predicting soil properties using models trained on soil spectral datasets is an important component of large scale projects such as the RaCA project (Wills *et al.*, 2014; Wijewardane *et al.*, 2016) which collected around 144,000 samples from across the conterminous United States for carbon stock mapping using vis-NIR, or the LUCAS project (Stevens *et al.*, 2013) in Europe with around 20,000 topsoil observations for general soil assessment.

While we have big datasets and global projects, currently, modelling is mostly based on "traditional" ML approaches: tree-based models (Cubist, random forest) in DSM and PLSR in spectroscopy. These methods are useful in certain situations but have

limitations, especially when dealing with either spatial data or signals, since they do not provide an effective way of incorporating contextual information, ignoring the spatial structure (1-, 2- or 3-D) of the data. Additionally, these models, particularly PLSR, are not suited for very large datasets which does not allow us to take full advantage of these data compilations.

In the context of DSM, large extent mapping mainly relies on datasets from different parties, which are then compiled into a single source and used by one person or organisation to train models. This workflow can be severely limited by privacy concerns, despite the positive effects of collaboration and data sharing (Fienberg *et al.*, 1985). While there are calls for "bottom up" approaches (e.g. GlobalSoilMap), they have not been generally applied and there are technical challenges such as harmonisation to remove the "patchwork" effect caused by using maps from different sources. All the above, in addition to the performance differences derived from the generality of global models versus the specificity of local models, there is still a debate on which approach is better.

Either in global or local DSM, most applications are static, ignoring the dynamism of soil properties such as soil organic carbon (SOC). Even when studies call themselves dynamic, it usually means that they change one or two factors (land use, climate) but model each situation without considering time. A better term, perhaps, is "partially dynamic soil scenario maps", as per McBratney *et al.*, (2003). Even with the large soil dataset available to date, in most situations, there is not enough soil data to perform a fully-empirical, space-time modelling under the DSM framework.

The application of ML in soil sciences usually prioritises numerical or categorical data over qualitative descriptions, which are usually considered subjective in nature (McBratney and Odeh, 1997). However, it must be taken into account the resources that have been invested in collecting large amounts of descriptive information. Neglecting descriptive data due to its inconsistency seems wasteful, yet methods that take advantage of this data have rarely been applied in soil sciences.

Considering these limitations, it is important to evaluate what advances in ML can be implemented as potential solutions. I propose the use of the following methods to solve the aforementioned limitations:

- The use of deep learning models, specifically convolutional neural networks

(CNNs) to incorporate contextual information, predict multiple properties simultaneously and make better use of large soil spectroscopy databases.

- Using online-learning methods for collaborative modelling while preserving data privacy, allowing the model to learn from different datasets that do not need to be compiled into a single source.
- The use of transfer learning to train a CNN to use knowledge extracted from a global dataset and apply it into a local context.
- Combine mechanistic models and ML methods in the context of space-time modelling of soil organic carbon.
- Use natural language processing (NLP) to incorporate descriptive data into numerical analyses.

The following sections expand on these points, implemented during this thesis as a solution to the aforementioned limitations.

Incorporating contextual spatial information

Digital soil mapping aims to model the relationships between soil attributes or classes and soil forming factors (McBratney *et al.*, 2003). The whole framework is based on digital information where soils are represented as points or polygons, and soil forming factors or their proxies as spatially continuous grids of environmental covariates. In the case of point observations, the corresponding soil forming factors are extracted from the pixels intercepted by the observations. This approach has been widely used and it is the base of most current digital soil maps. A limitation of this approach is that, by only extracting the covariate information that matches the coordinates of the observations, we are completely ignoring the landscape where the observations are embedded.

Some articles have tried to propose solutions to include spatial context. Most of these works pre-calculate covariates using a window around the observations, at different scales, to then feed them to models such as random forest or fully-connected neural networks (e.g. Behrens *et al.*, (2018)). The problem with that approach is

two-fold: a) the models are only capable of handling 1-dimensional arrays, ignoring the inherent spatial structure of the data; and b) the pre-calculation or transformation, commonly known as feature engineering, is time-consuming since it heavily relies on manual work and it is arbitrarily driven by domain knowledge (Khurana *et al.*, 2016).

In this work we propose the use of CNNs to solve both problems. First, CNNs are very flexible and are capable of dealing with higher-dimensional data. In a 2-dimensional case, we are able to use the same window around the observations and, thanks to the convolution operations, the data is processed, preserving its spatial structure. Second, CNNs have the capacity to automatically create new features as the data passes through the network, relieving researchers from the burden of manual feature engineering (Barz *et al.*, 2017), and also exploring features beyond their domain knowledge. By incorporating contextual spatial information we expect an improvement of the model's performance and more realistic output maps.

Simultaneous prediction of multiple properties

Most modelling applications in soil science do not consider the relation between soil properties to improve their generalisation power. It is easy to think about some examples where predicting multiple properties could be beneficial. Bulk density (Bd) is notoriously difficult to predict using Vis-NIR spectroscopy, whereas SOC is one of the properties routinely predicted with good performance. The relation between SOC and Bd has been reported since early years of soil science (e.g. Curtis and Post (1964)) and, ideally, we would like to exploit that relationship by using the generalisation learned to predict SOC as a guide when learning the generalisation for Bd. Similarly, when we consider soil properties at different depths, our model should be able to understand that they are related, similar to a depth function representation (Nakane, 1976). Sequeira *et al.*, (2014) predicted Bd to populate an incomplete database and concluded that information about the Bd of other horizons in the soil profile was the most important variable. Despite the evident advantage of using auxiliary information, most of the DSM studies treat the prediction of soil properties using more or less depth-independent layers (Arrouays *et al.*, 2017).

Thanks to the flexibility of neural networks, in this work we explore the use of multi-task learning to train a model aiming to simultaneously predict multiple

properties. The model should be able to learn representations common for all the properties, representations per individual property and also how they are all related to each other in order to boost its generalisation power. We apply this concept to soil spectroscopy data and also in the context of DSM, where multiple properties and multiple depths are simultaneously predicted, respectively. An increase in the generalisation power is expected, which is translated into more accurate predictions.

Global and local models

Many initiatives are focused on collating or generating new soil datasets that cover large extents. As a consequence, we have seen the development of many models with global or continental coverage. Many studies have evaluated the performance of those models at local scale, comparing them with models generated with local datasets, generally concluding that the local models outperform the more general model. For instance, Mulder *et al.*, (2016) and Griffiths *et al.*, (2016) reported that a global map was a poor representation compared to national maps, for SOC in France and pH in England, respectively. This is an expected outcome since global models are usually exposed to more variety of soils, generalising over a broader range of values for a soil property.

It seems that "local" and "global" are in two opposite sides, and that having exposure to more varieties of soil is counter-productive when we consider a local application. Two of the most widely used methods to overcome this problem are spiking and sub-setting. Spiking consists of adding a small number of local samples to the larger global dataset followed by a re-calibration of the model (Shepherd and Walsh, 2002). Sub-setting consists fn selecting a subset of the global dataset that resembles the local data based on some measures of similarity (Araújo *et al.*, 2014), whether spectral similarity or geographical proximity. The commonality between both methods is that they ignore valuable global information. What if we could use that "experience" obtained from the extensive variety of soils when predicting at local scales? The process of sharing representations between different domains is known as transfer learning (Pan and Yang, 2010). This research explores the use of transfer learning to "localise" a general continental model, trained on a spectral soil library. It is expected that some knowledge extracted by the general model can be used when training the local models, yielding some improvement over model generated only using the local

data.

Privacy-preserving collaborative modelling

Collaboration is an important aspect of research, specially when trying to tackle large scale problems. Data collection to solve such problems is impractical and sharing data between research groups becomes a necessity. Despite the positive effects of collaboration and data sharing, the latter comes with concerns about privacy and confidentiality (Fienberg *et al.*, 1985).

To overcome this challenge, we propose the novel use of an advanced ML training technique, namely online learning. The technique is designed to handle big datasets that do not fit in the user's computer memory or streams of data where new data is constantly generated (Toledo, 1999; Lee *et al.*, 2010). Both situations imply that the complete dataset is not available during training. This concept is easily expanded to a situation when the data is held by different parties, at different locations. Instead of moving data between different collaborators, this method moves the model from one location to the next, effectively transferring the "knowledge" incrementally obtained from the different datasets. It is expected that the model performance will not be significantly different from a model trained using all the data at once.

Global space-time mapping using semi-mechanistic models

The number of soil profiles available to date, in a harmonised global database, falls between 100,000 and 200,000. Considering all the possible combinations between the different soil formation factors, including for example temperature, precipitation and the land cover dynamics, this number of samples is not enough to fully exploit the capacity of advanced machine learning models. An alternative approach is to use machine learning as a component of a more complex, mechanistic model. This semi-mechanistic approach has the advantage of incorporating into the model knowledge acquired *a-priori* (Braake *et al.*, 1999).

Using a global, spatiotemporal assessment of SOC as an example, I present a semi-mechanistic model that incorporates DSM with time-for-space substitution (similar to the space-for-time approach commonly used in ecology as an alternative to long-term studies (Pickett, 1989)) and a land cover tracking component which

takes into account landcover and SOC dynamics. Thanks to the landcover tracking component, the resulting model has a very important quality — memory — effectively incorporating the frequency, sequence, time span, and magnitude of changes which are vital to understand the impact of human activity on the world's ecosystems (Watson *et al.*, 2014).

Using descriptive data

Soil data collection is a multi-step process including field, laboratory and "desk" analyses. Throughout this process, a variety of data is obtained and, depending on the analysis, its nature differs significantly. For instance, most of the traditional data collected in the field are of descriptive nature, perhaps subject to bias, but giving important contextual information about the location of an observation. On the other hand, data obtained by wet chemistry in the laboratory are usually considered as highly precise measurement. Due to the lower uncertainty of laboratory data and the fact that modelling methods usually favour numerical data, descriptive field data has been greatly neglected.

In this thesis, I proposed the use of NLP, which involve the manipulation and analysis of language (Jain *et al.*, 2018), to extract information from descriptive data, specifically field descriptions. Thanks to NLP, it is possible to generate a multi-dimensional vector space where words' locations depend on their syntactic and semantic relationships. By using this method, we are effectively representing words as numbers, which allow us to incorporate this data into traditional numerical analysis.

References

Araújo, S., Wetterlind, J, Demattê, J., and Stenberg, B (2014). Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *European Journal of Soil Science* 65 (5): 718–729.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., *et al.*, (2014). GlobalSoilMap: Toward a fine-resolution global grid of soil properties. In: *Advances in agronomy*. Vol. 125. Elsevier: pp. 93–134.

Arrouays, D., Lagacherie, P., and Hartemink, A. E. (2017). *Digital soil mapping across the globe*.

Barz, B., Rodner, E., Käding, C., and Denzler, J. (2017). Fast Learning and Prediction for Object Detection using Whitened CNN Features. *arXiv preprint arXiv:1704.02930*.

Behrens, T., Schmidt, K., MacMillan, R. A., and Rossel, R. A. V. (2018). Multi-scale digital soil mapping with deep learning. *Scientific reports* 8.

Braake, H. T., Roubos, J., and Babuška, R (1999). Semi-mechanistic modeling and its application to biochemical processes. In: *Fuzzy Logic Control: Advances in Applications*. World Scientific: pp. 205–226.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Briggs, L. J. and McLane, J. (1907). The moisture equivalent of soil. *USDA Bureau of Soils Bulletin* (45): 1–23.

Briggs, L. J. and Shantz, H. (1912). The wilting coefficient and its indirect determination. *Botanical Gazette* 53 (1): 20–37.

Burrough, P. A., Bouma, J., and Yates, S. R. (1994). The state of the art in pedometrics. *Geoderma* 62 (1-3): 311–326.

Curtis, R. O. and Post, B. W. (1964). Estimating Bulk Density from Organic-Matter Content in Some Vermont Forest Soils 1. *Soil Science Society of America Journal* 28 (2): 285–286.

Dale, M., McBratney, A., and Russell, J. (1989). On the role of expert systems and numerical taxonomy in soil classification. *Journal of Soil Science* 40 (2): 223–234.

Fienberg, S. E., Martin, M. E., Straf, M. L., Council, N. R., *et al.*, (1985). *Sharing research data*. National Academies.

Griffiths, R. I., Thomson, B. C., Plassart, P., Gweon, H. S., Stone, D., Creamer, R. E., Lemanceau, P., and Bailey, M. J. (2016). Mapping and validating predictions of soil bacterial biodiversity using European and national scale datasets. *Applied soil ecology* 97: 61–68.

Jain, A., Kulkarni, G., and Shah, V. (2018). Natural language processing. *International Journal of Computer Sciences and Engineering* 6 (1): 161–167.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science* 349 (6245): 255–260.

Khurana, U., Turaga, D., Samulowitz, H., and Parthasrathy, S. (2016). Cognito: Automated feature engineering for supervised learning. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE: pp. 1304–1307.

Lee, M., hong Jeon, J., Kim, J., and Song, J. (2010). Scalable and parallel implementation of a financial application on a GPU: With focus on out-of-core case. In: *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*. IEEE: pp. 1323–1327.

McBratney, A., Mendonça Santos, M. L., and Minasny, B (2003). On digital soil mapping. *Geoderma* 117 (1): 3–52.

McBratney, A. B. and Odeh, I. O. (1997). Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma* 77 (2-4): 85–113.

Minasny, B., Malone, B. P., Mcbratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B. S., *et al.*, (2017). Soil carbon 4 per mille. *Geoderma* 292: 59–86.

Mulder, V., Lacoste, M, Richer-de Forges, A., Martin, M., and Arrouays, D (2016). National versus global modelling the 3D distribution of soil organic carbon in mainland France. *Geoderma* 263: 16–34.

Nakane, K. (1976). An empirical formulation of the vertical distribution of carbon concentration in forest soils. *Japanese Journal of Ecology* 26 (3): 171–174.

Pan, S. J., Yang, Q., *et al.*, (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22 (10): 1345–1359.

Pickett, S. T. (1989). Space-for-time substitution as an alternative to long-term studies. In: *Long-term studies in ecology*. Springer: pp. 110–135.

Sequeira, C. H., Wills, S. A., Seybold, C. A., and West, L. T. (2014). Predicting soil bulk density for incomplete databases. *Geoderma* 213: 64–73.

Shepherd, K. D. and Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil science society of America journal* 66 (3): 988–998.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., and van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PloS one* 8 (6): e66409.

Toledo, S. (1999). A survey of out-of-core algorithms in numerical linear algebra. *External Memory Algorithms and Visualization* 50: 161–179.

Watson, S. J., Luck, G. W., Spooner, P. G., and Watson, D. M. (2014). Land-use change: incorporating the frequency, sequence, time span, and magnitude of changes into ecological research. *Frontiers in Ecology and the Environment* 12 (4): 241–249.

Wijewardane, N. K., Ge, Y., Wills, S., and Loecke, T. (2016). Prediction of soil carbon in the conterminous United States: Visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project. *Soil Science Society of America Journal* 80 (4): 973–982.

Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., and West, L. T. (2014). Overview of the US rapid carbon assessment project: sampling design, initial summary and uncertainty estimates. In: *Soil carbon*. Springer: pp. 95–104.

Chapter 1

Machine learning and soil sciences: A review aided by machine learning tools

Summary

The application of machine learning (ML) techniques in various fields of science has increased rapidly, especially in the last ten years. The increasing availability of soil data that can be efficiently acquired remotely and proximally, and freely available open-source algorithms, have led to an accelerated adoption of ML techniques to analyse soil data. Given the large number of publications, it is difficult to manually review all papers on the application of ML in soil science without narrowing down a narrative of ML application to a specific research question. This paper aims to provide a comprehensive review of the application of ML techniques in soil science aided by a ML algorithm (Latent Dirichlet Allocation) to find patterns in a large collection of text corpus. The objective is to gain insight into publications of ML applications in soil science and to discuss the research gaps in this topic. We found that: a) there is an increasing usage of ML methods in soil sciences, mostly concentrated in developed countries, b) the reviewed publication can be grouped into 12 topics, namely remote sensing, soil organic carbon, water, contamination, methods (ensembles), erosion and parent material, methods (NN, SVM), spectroscopy, modelling (classes), crops, physical and

modelling (continuous), c) advanced ML methods usually perform better than simpler approaches thanks to their capability to capture non-linear relationships. From these findings, we found research gaps, in particular: about the precautions that should be taken (parsimony) to avoid overfitting, and that the interpretability of the ML models is an important aspect to consider when applying advanced ML methods in order to improve our knowledge and understanding of soil. We foresee that a large number of studies will focus on the latter topic.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for this publication is:

Padarian, J., Minasny, B., and McBratney, A. B. 2020. Machine learning and soil sciences: a review aided by machine learning tools, Soil, 6, 35–52.

Contributors	Statement of contribution
José Padarian *Signature:* José Padarian *Date:* April 2, 2020	Conceptualisation Data analysis Writing
Budiman Minasny	Writing
Alex McBratney	Writing

The application of machine learning (ML) techniques in various fields of science has increased rapidly, especially in the last ten years. Soil science research, in particular, Pedometrics, has used statistical models to "learn" or understand from data how soil is distributed in space and time (McBratney *et al.*, 2019). The increasing availability of soil data that can be efficiently acquired remotely and proximally, and freely available open-source algorithms, have led to an accelerated adoption of ML techniques to analyse soil data. Several well-known ML applications in soil science include the prediction of soil types and properties via digital soil mapping (DSM) or pedotransfer functions, and analysis of infrared spectral data to infer soil properties. Machine learning analysis of soil data is also used to draw conclusions on the controls of the distribution of the soil.

The definition or what constitutes ML is still contentious. In this work, instead of adding a new argument to differentiate ML from statistical science, we will focus on the view of Jordan and Mitchell (2015) where ML is "lying at the intersection of computer science and statistics". With respect to artificial intelligence (AI), sometimes we have seen the terms ML and AI used interchangeably. This is understandable confusion since ML is a subset of AI, but not everything related to AI falls in the ML category (e.g. expert systems).

There are concerns that ML applications ignore soil science knowledge (Rossiter, 2018), and that the results could be misleading and wrong. Nevertheless, many would find that ML methods can help in the scientific process (Mjolsness and DeCoste, 2001; Rudin and Wagstaff, 2014): observations, empirical and theory-based models development, and simulations of soil processes (Rossiter, 2018). For example, exploration of high-dimensional infrared spectral data helps in understanding the horizonation designation in a soil profile (Fajardo *et al.*, 2016). The process of modelling and validation can be used to formulate a model to explain soil distribution (Brungard *et al.*, 2015). Modelling via ML can also be used to improve our understanding of the causes of soil variation. Results from ML models can inform which environmental variables control soil distribution. New relationships revealed by ML analysis can help to stimulate ideas, generate hypotheses, and formulate future questions for research (Ma *et al.*, 2019).

This paper aims to provide a comprehensive review of the application of ML techniques in soil science. A quick Google Scholar search of "soil" and "machine

learning" resulted in more than 70,000 items, with 16,000 items published in 2018. While we can narrow down a narrative of ML application to a specific research question, such as the application of ML to yield prediction in precision agriculture (Chlingaryan *et al.*, 2018) or DSM, it is difficult to manually review all papers on the application of ML in soil science. One ML technique that has not been applied in soil science is topic modelling, a type of quantitative text mining method. Similar to what ML does to numerical data, topic modelling finds patterns in a large collection of text corpus (Blei *et al.*, 2003; Blei, 2012) and it has been used to study the evolution of various disciplines and topics (Zhou *et al.*, 2006; Sugimoto *et al.*, 2011; Wu *et al.*, 2014).

This paper uses topic modelling to analyse the trend in ML application in soil science. The objective is to gain insight into publications of ML applications in soil science, in particular, we will try to answer the following questions:

- Who is using ML?, Is the application of ML as ubiquitous as we think? and
- Which ML methods are commonly used and how often have they been used?
- In which areas of soil sciences do we use ML? and how are they clustered and related?
- Do advanced ML methods perform significantly better than linear or non-linear statistical approaches?
- Can ML methods simulate soil processes in space and time?
- Can we use ML methods to improve our knowledge and understanding of soil?

Throughout this review, we will refer to models as "simple" or "complex/advanced" trusting in the readers' criteria. To illustrate that gradient between simple and complex, we considered a linear model (LM) with two variables as simple compared to a LM with 100 variables; a classification and regression tree (CART) with two branches as simple compared to a CART with 100 branches; and finally, a CART with two branches as simple compared with a LM with 100 variables. We also hope that it is clear for the reader that a model such as a deep convolutional neural network (CNN) has many parameters, hence is more complex than a CART model.

1.1 Methods

1.1.1 Article selection

In order to identify the primary group of articles, we used the term "soil 'machine learning'" to perform a full-text search in databases from different publishers. We selected the publishers based on a) our institution having access to full-text articles, and b) that they provide text-mining permission. We limited our search to English only literature, without fixing a specific time-frame, and completing the search the 1st of February 2019. After performing a screening for the relevance of the initial 3044 matches, we decided to narrow down the selection to the articles containing the word "soil" in their title, yielding a total of 322 articles. The final journal names and number of articles are shown in Table 1.A.1.

1.1.2 Topic modelling

Topic modelling is a probabilistic ML method that aims to discover and annotate large archives of documents with thematic information (Blei, 2012). By analysing the words contained in a set of documents, these topic modelling algorithms are capable of identifying common themes. These methods allow processing an arbitrarily large number of articles, which can help to reduce part of the bias introduced by only selecting a manageable subset of documents, or by manually assigning documents to topics.

In order to determine in which areas of soil sciences we use ML, we selected an algorithm commonly used in topic modelling called Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) to perform the task of allocating the articles into topics. LDA is a probabilistic model that assumes that a number of topics exist in a document collection and each topic is represented by a distribution of words. Each document is represented by a distribution over topics, and each word is a sample over each topic's vocabulary (Fig. 1.1). For more details about the LDA, we refer the reader to Blei (2012).

Before modelling the topics, we pre-processed the documents in order to reduce the noise of the unstructured texts. We a) removed stop-words (common words such as "from" and "are"), b) we generated bi- and tri-grams, which are groups of 2 and 3 words which commonly appear together in the text (e.g. "remote sensing", "particle size distribution"), and c) removed extremely uncommon (that appear in less than 5

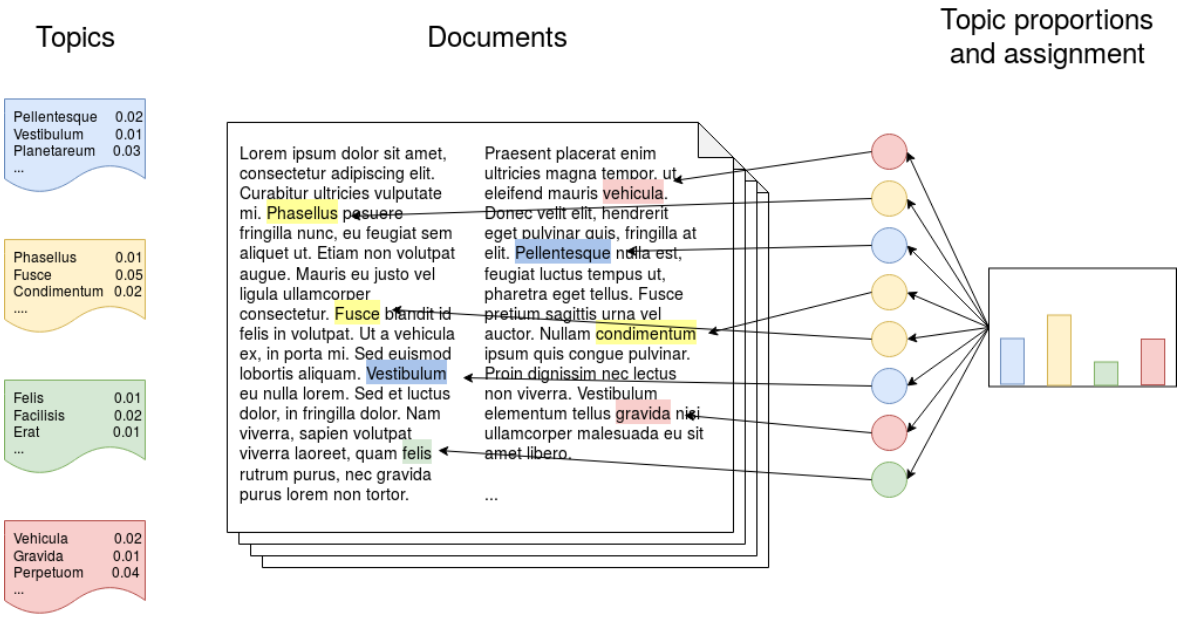


Fig. 1.1: Collection of topics with distribution of words (left), document distribution over topics (histogram, right) and words sampled from the topics' vocabularies (circles). The topics, words and assignment are for illustrative purposes. Adapted from Blei (2012).

documents) and common words (that appear in more than 50% of the documents), which do not help to differentiate between topics.

The LDA algorithm is capable of learning different topics to which each document is assigned given the words that constitute it. The first challenge is to find the optimal number of topics, which has to be general enough to capture similarities between articles but with some degree of specificity in order to have a manageable and sensible group of topics. That balance between generality and specificity is key to generate topics that are semantically interpretable by a human (Stevens *et al.*, 2012). One of the measures that is highly correlated with human interpretation of the topics is topic coherence (Stevens *et al.*, 2012). We estimated a coherence measure proposed by Röder *et al.*, (2015) (referenced as *CV* in their paper) for different models trained with an increasing number of topic, from 2 to 30. *CV* is an aggregated measure which combines a normalised point-wise mutual information coherence measure, cosine vector similarity and a boolean sliding window of size 110. It ranges from 0 to 1, with 1 being the highest coherence. Other parameters of the LDA algorithm such as the threshold of

the probability above which a topic considered, and the number of training iterations were set to 0.2 and 1000, respectively, after performing a parameter grid search.

1.1.3 Text extraction

In order to identify the information required to answer our questions, we used a combination of named-entity recognition and rule-based matching. To extract the MODEL entities, we used a list of modelling methods from the *Outline_of_machine_learning* Wikipedia article in addition to other algorithms that are commonly used in soil sciences and that were not present in the list (e.g. Cubist). After extracting the MODEL entities, we proceeded to extract the abbreviations used to reference those models. In order to extract the abbreviations, we relied on the commonly seen pattern of writing model names followed by their corresponding abbreviation (e.g. "we used a random forest model (RF)"). By extracting the abbreviations we expected to discriminate between a) models used to generate the results reported in the articles and b) models mentioned to give context to the studies. Extracting abbreviations also allowed us to capture variations of models not present in our original list (e.g. BART for bagged regression trees in Fig. 1.2). It is important to note that throughout this work we used text extraction just to aid the reviewing process. All the evaluations were semi-manual. First, we identified articles mentioning specific keywords to then manually inspect the selected articles.

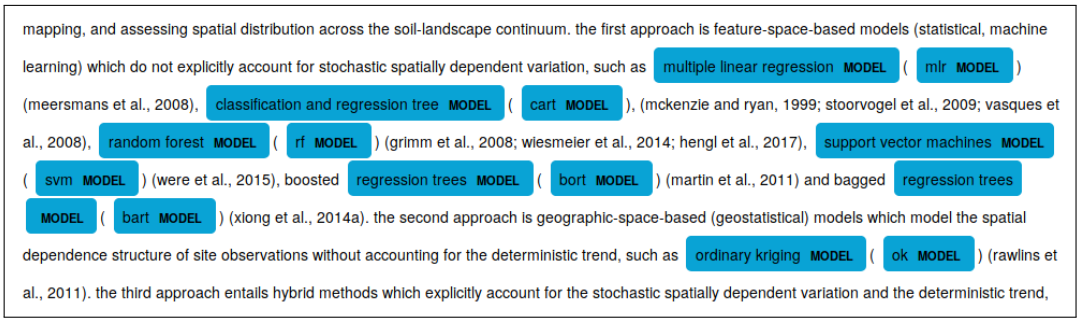


Fig. 1.2: Excerpt from one of the reviewed articles showing named entities recognised as models. Note that the word 'bagged' is not recognised, but the abbreviation 'bart' is.

1.1.4 Implementation

We performed all our analysis in Python, using the libraries gensim v3.6.0 (Řehůřek and Sojka, 2010) and the MALLET package (McCallum, 2002) for the topic modelling and spacy v2.1.0a6 (Matthew and Honnibal, 2017) for the named entity recognition.

1.2 Results and discussion

1.2.1 Who is using machine learning methods?

The first questions related to the current status of the ML literature in soil sciences can be answered after correctly organising all the articles' metadata. Regarding the general usage of ML methods, in our review, we observed an expected increment in time in the number of publications using ML to model different aspects of soils (Fig. 1.3). This increment is most likely due to a combination of increasing computational power and accessibility to high-performance computers, increasing availability of data (e.g. remote sensing) (Jordan and Mitchell, 2015), and the increasing interest in "data science". It is also confounded with the overall increase in the number of publications, which was estimated in 2015 at nearly 2.5 million new publications per year (Ware and Mabe, 2015).

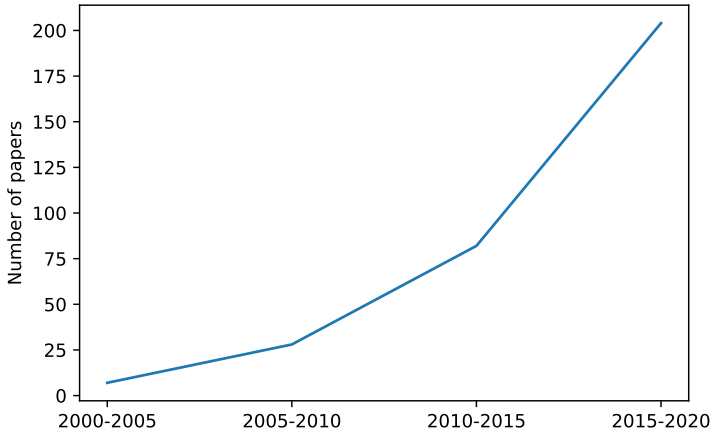


Fig. 1.3: Distribution in time of the articles used in this review.

Besides the temporal trend in publishing, we were also interested in how ubiquitous the application of ML methods is. Fig. 1.4 shows the number of institutions per country

($\log_{10}$) that appeared listed as an affiliation in the analysed articles. ML techniques in the context of soil sciences are used by authors from many countries around the world but mostly concentrated in developed countries. This is due to the inseparable relationship between science, technology and development (Sagasti, 1973), which is also related to what is usually called "digital divide" (Rossiter, 2018). Inter-institutional collaboration could be an important aspect of closing this gap (Sonnenwald, 2007). Similar to what is happening in many disciplines (Sonnenwald, 2007), we observed an increase in the number of co-authors per article (Fig. 1.5), which might be a good sign if we avoid bad practices like "helicopter science" (Minasny and Flantis, 2018).

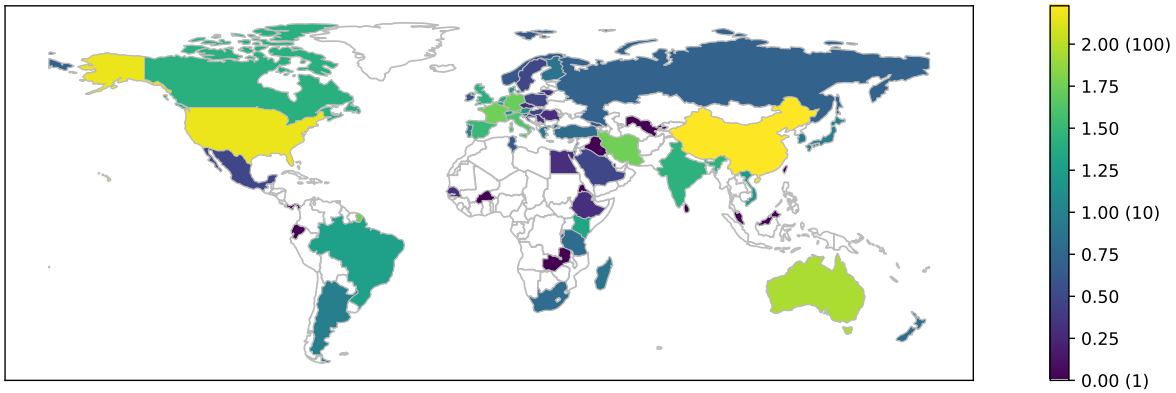


Fig. 1.4: Total number ($\log_{10}$) of institutions per country that participated in articles included in this review. Numbers between brackets are real number of publications. Outlined countries have zero occurrences.

The advance of a discipline is not only measured by the number of publications. Dissemination of knowledge is a key component of research and Open Access (OA) has been recognised as an optimal solution since it is in the best interests of all stakeholders involved in the process (Björk, 2017). In the application of ML in soil sciences, the proportion of OA publications is very low (Fig. 1.6). This number is in line with the overall OA presence in science (Björk, 2017) but on the opposite side of the general trend in ML literature where scientists prefer AO (Hutson, 2018).

1.2.2 Most used methods

From the huge variety of ML models available, we found over 100 different variants that have been applied in soil sciences. From those, most have been applied experimentally

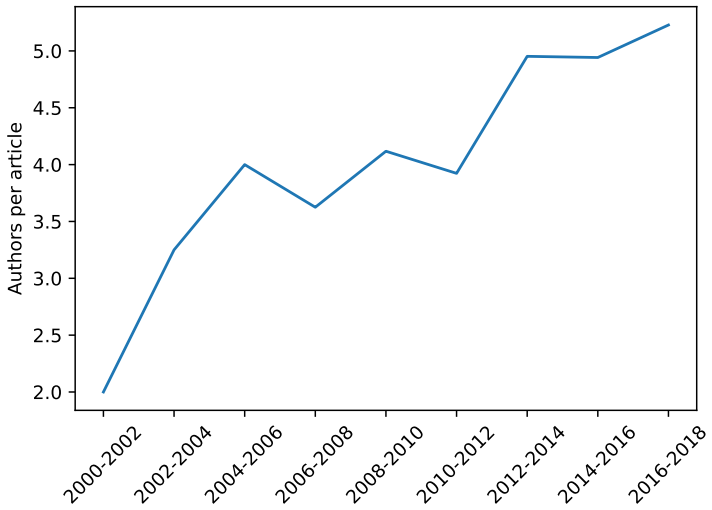


Fig. 1.5: Evolution of the number of authors per publication. Mean values per time-period.

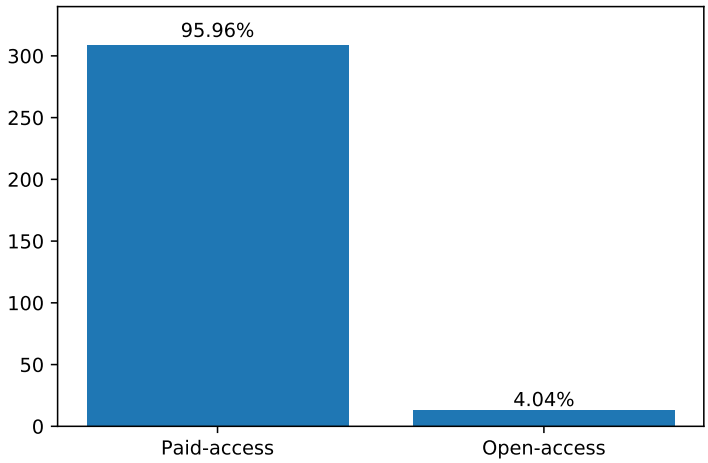


Fig. 1.6: Number of articles published under paid and open access.

in one or two papers and just a hand-full are consistently used. Fig. 1.7 depicts the evolution of some selected models. There is an overall increase in the usage of all the models but, proportionally, it is possible to see a decrease in the usage of some models such as support vector machines (SVM), Multivariate Adaptive Regression Spline (MARS) and CART, giving the way to more advanced alternatives such as random forest (RF). The adoption of the latter has an accelerated growth and it has been used in a diversity of topics, including mapping and spectroscopy. It is also

noticeable the appearance of deep learning, which at the moment has only been used in a few publications related to mapping and spectroscopy.

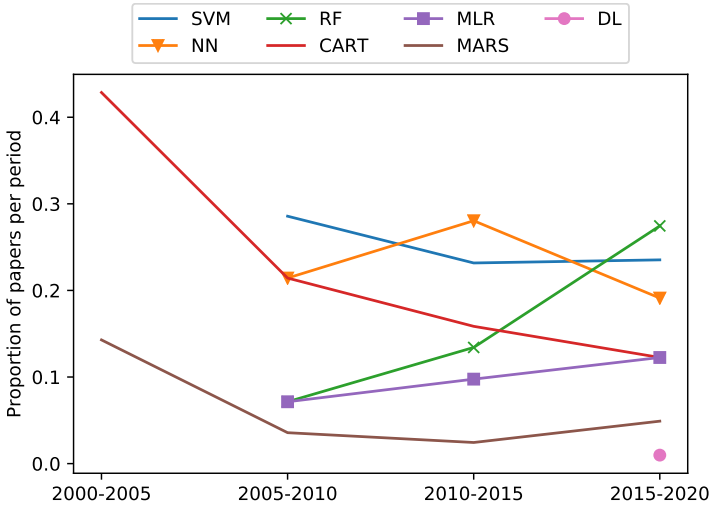


Fig. 1.7: Evolution of model usage in time. SVM: support vector machines; NN: neural networks; RF: random forest; CART: classification and regression trees; MLR: multiple linear regression; MARS: multivariate adaptive regression spline; DL: deep learning

.

1.2.3 Main topics

As we mentioned in Section 1.1.2, in order to find the optimal number of topics present in the corpora, we trained models with an increasing number of topics (from 2 to 30) and we plotted the evolution of the *CV* coherence (Fig. 1.8). From this curve it is possible to select the number of topics that yield the highest coherence, which in this case is 12.

These 12 topics correspond to main soil areas detected by the LDA algorithm where ML is applied. We extracted the most relevant words for each of the 12 topics and we examined the titles of the more relevant papers to identify suitable "topic names". The 12 identified areas were:

Remote sensing: Articles heavily based on remote sensing (Grunwald *et al.*, 2015; Xu *et al.*, 2017; Zhang *et al.*, 2018b). Articles related to salinity were also assigned to this group since most of them use remote sensing techniques (Khadim *et al.*, 2019; Zhang *et al.*, 2019).

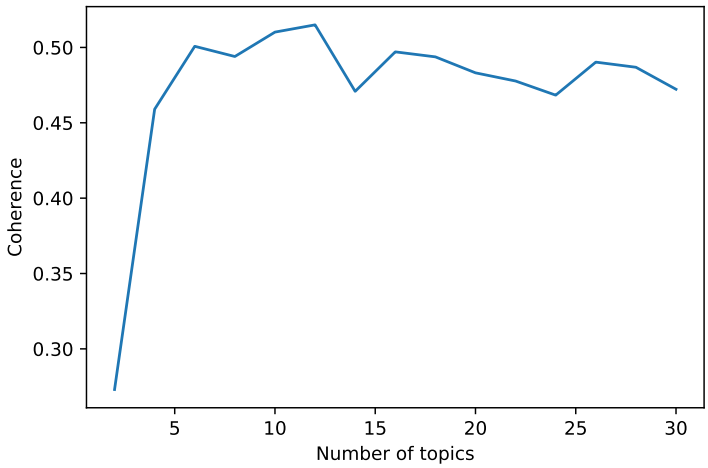


Fig. 1.8: Coherence by number of topics used to train a LDA model.

Soil organic carbon: Articles related to soil organic carbon (SOC) cycles and dynamics, and its relationship with the environment. Carbon stocks in different ecosystems, with particular emphasis in grasslands and topsoil (Rial *et al.*, 2017; Liu *et al.*, 2018; Song *et al.*, 2018; Wang *et al.*, 2018a).

Water: Articles mostly focused on soil water content and its changes over time (Ahmad *et al.*, 2010; Coopersmith *et al.*, 2014; Greifeneder *et al.*, 2018; Han *et al.*, 2018). Other articles in this category are related to soil temperature and CO_2 fluxes (Xing *et al.*, 2018; Oh *et al.*, 2019; Warner *et al.*, 2019; Zeynoddin *et al.*, 2019). All these articles comprise measurements made by "stations".

Contamination: Articles addressing problems related to heavy metals, soil pollution and bio-availability (Costa *et al.*, 2017; Reeves *et al.*, 2018; Wu *et al.*, 2013).

Methods (ensembles): Articles with a focus on model ensembles such as RF (Blanco *et al.*, 2018; Tziachris *et al.*, 2019).

Erosion / parent material: Articles focused on soil formation processes, specifically additions and losses by deposition and erosion, respectively (Geissen *et al.*, 2007; Märker *et al.*, 2011; Martinez *et al.*, 2017). Since soil formation depends on the parent material, articles aiming to characterise it were also included in this category (Kheir *et al.*, 2008; Lacoste *et al.*, 2011).

Methods (NN, SVM): Articles with a focus on methods such as neural network (NN) and SVM (Kovačević *et al.*, 2010; Farfani *et al.*, 2015; Hanna *et al.*, 2007).

Spectroscopy: This topic is related to proximal soil sensing covering different sections of the electromagnetic spectrum, from microwave to infrared to gamma (Heggemann *et al.*, 2017; Butler *et al.*, 2018; Xie and Li, 2018).

Modelling (classes): Articles focused on the modelling, especially mapping, of categorical soil properties based on their relationship with environmental covariates (Mansuy *et al.*, 2014; Camera *et al.*, 2017; Dharumarajan *et al.*, 2017; Massawe *et al.*, 2018). In this category is also possible to find articles related to the use of conventional soil maps, especially spatial disaggregation of polygons (Subburayalu *et al.*, 2014; Vincent *et al.*, 2018; Flynn *et al.*, 2019).

Crops: This group of articles focused not merely on soil but on its interaction within the soil-plant continuum. Water and nutrient availability in order to assure crop yields is a key component of this topic (Karandish and Šimůnek, 2016; Ivushkin *et al.*, 2018; Khanal *et al.*, 2018; Leenaars *et al.*, 2018).

Physical: Articles related to the physical properties of soils, including texture and bulk density (Bondi *et al.*, 2018; Naderi-Boldaji *et al.*, 2019), and how they affect aspects of soil such as water retention and flow (Koestel and Jorda, 2014; Gao *et al.*, 2018).

Modelling (continuous): Articles focused on the modelling, especially mapping, of continuous soil properties based on their relationship with environmental covariates, from regional to continental scales (Henderson *et al.*, 2005; Dai *et al.*, 2014; Poggio *et al.*, 2016; Padarian *et al.*, 2019a; Caubet *et al.*, 2019). In this category it is also possible to find articles related to pedotransfer functions (Dobarco *et al.*, 2019).

These topics are not completely independent and they share some commonalities. For instance, Fig. 1.9 shows an overlap between Topic 12 (Modelling continuous properties) and 9 (Modelling classes) since both are related to mapping using environmental covariables. Both topics are also related to topic 3 (Water) since its

articles usually have a spatial component. Something similar occurs between Topic 8 (Spectroscopy) and 1 (Remote sensing) since both are related to spectral data.

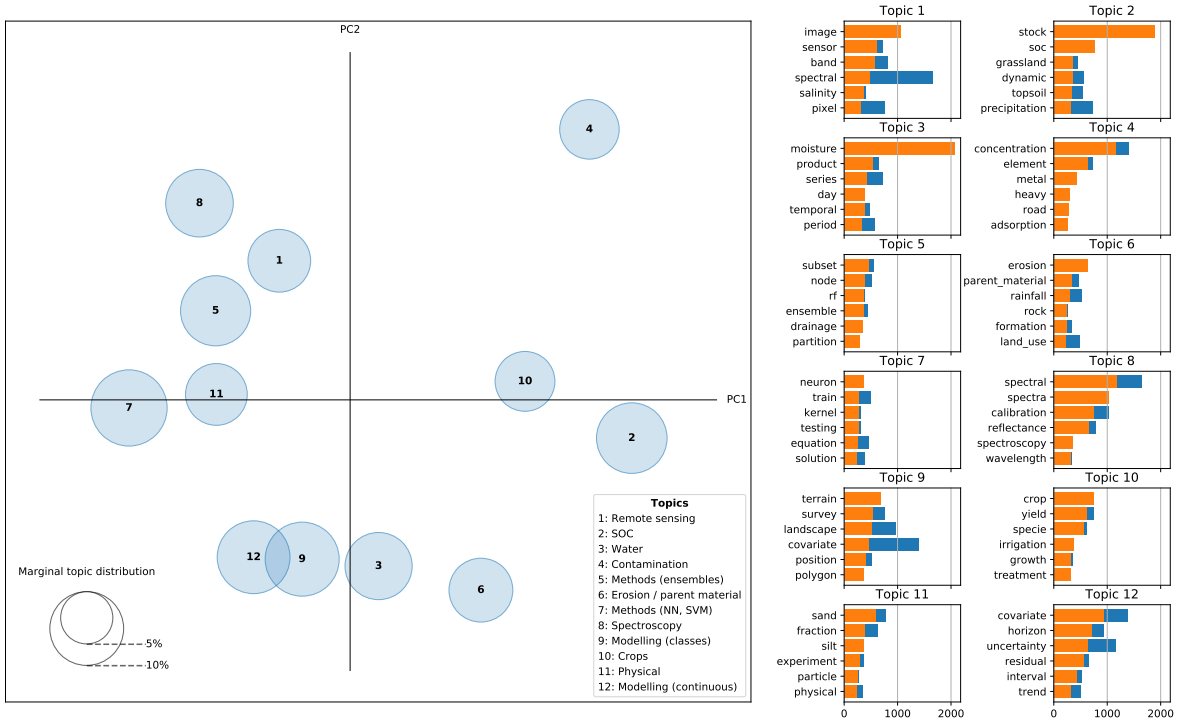


Fig. 1.9: Inter-topic Distance Map. Dimension reduction via Jensen-Shannon Divergence (Lin, 1991) to measure the distance between probability distribution of words and Principal Coordinate Analysis. Top-6 more relevant words per topic. Complete bars (blue + orange) correspond to overall term frequency and shaded (orange) correspond to term frequency within the selected topic.

Besides the shared features between topics, given that LDA is a probabilistic model, articles also contain features related to more than one topic, i.e. they talk about more than one topic (Fig. 1.10). For instance, many of the articles related to SOC are also related to soil modelling and mapping (Deng *et al.*, 2018; Wang *et al.*, 2018b; Gomes *et al.*, 2019; Keskin *et al.*, 2019).

1.2.4 Performance of machine learning models

Our review shows that more advanced modelling techniques usually yield better results compared with simpler approaches. In one of the more extensive comparisons,

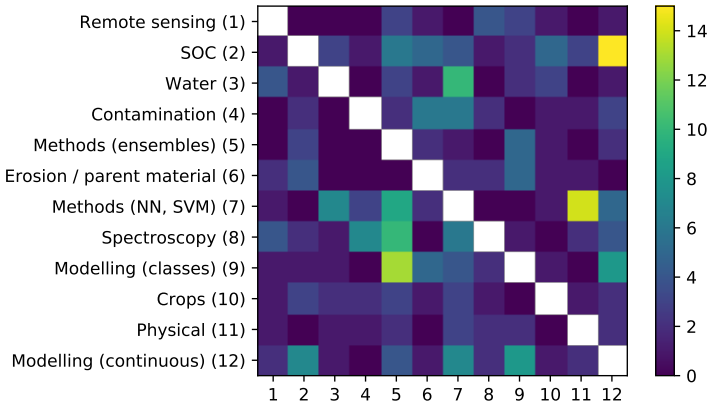


Fig. 1.10: Co-occurrence between the two most likely topics per document. Values correspond to number of papers.

Sirsat *et al.*, (2018) compared 76 different algorithms, where ensembles of extremely randomised regression trees ranked first when predicting soil fertility indices. Other comparative studies also showed a consistent higher performance of ML methods (NN, SVM, RF) over simpler approaches (Principal Components Regression, Partial Least Squares Regression (PLSR), multiple linear regression (MLR), k-Nearest Neighbours) in applications such as spectroscopy (Viscarra-Rossel and Behrens, 2010; Morellos *et al.*, 2016) and DSM (Brungard *et al.*, 2015; Taghizadeh-Mehrjardi *et al.*, 2015; Camera *et al.*, 2017; Jeong *et al.*, 2017). Most of these studies mention that the superiority of these algorithms is given by their capability to deal with complex nonlinearities present in the data. Moreover, the better performance of more advanced ML methods is reported in studies related to the prediction of continuous properties and classes.

Regarding the connection between performance and model usage (Section 1.2.2), we observed that some simpler methods such as MLR, despite their lower performance compared to more advanced models, are very popular. This is expected for statistical models since they have a long tradition in science. On the other hand, we also observed a natural tendency of leaving some model behind despite being used for a long time. For instance, PLSR is very popular and has been used since the 80-90s but, when used in the studies included in this review (mostly published post 2000s), very few studies use it as their main algorithm and, instead, it is used in comparative studies where it is outperformed by more advanced models.

It is worth noting that the final performance is not solely dependent on the selected

modelling method. Advanced methods like NNs have a big number of parameters to fit, especially in the context of deep learning. In order to correctly fit those parameters, from a computational and statistical point of view, the size of the dataset is an essential factor (Jordan and Mitchell, 2015). Padarian *et al.*, (2019b) show that a deep CNN trained using a large dataset (around 20,000 soil samples) outperformed methods such as PLS and Cubist when predicting soil properties from spectral data. Using the same method but training on a significantly smaller dataset (390 soil samples), the CNN yielded the worst results.

There is not a clear rule on how big a dataset should be, especially because it certainly depends on the complexity of the underlying problem, but the relationship between dataset size and performance has been shown in many studies, using what is usually known as "learning curves" (Catlett, 1991; Shavlik *et al.*, 1991; Cortes *et al.*, 1994; Perlich *et al.*, 2003; Somarathna *et al.*, 2017). During our review, we observed that the dataset size varied greatly depending on the ML methods (Fig. 1.11).

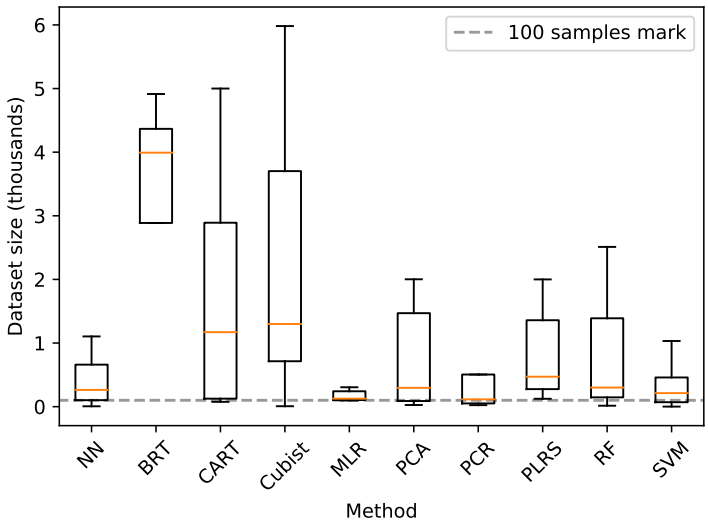


Fig. 1.11: Boxplot of reported dataset sizes grouped by method. Outliers were removed. The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR).

Considering that ML models could generate a similar solution to a linear model (e.g. a single-rule tree), it should not be a problem to use any method for any dataset size. However, the main difficulty is that training a complex ML model is not a trivial

task, especially to avoid overfitting and to obtain a good generalisation, which becomes challenging in the presence of small datasets (i.e. training and test datasets). Even if a researcher can overcome the training process, it is probable that a simpler model can yield similar results.

1.2.5 Space-time modelling

Compared with the spatial component of soil variation, which is prominent in the topics found using the LDA algorithm (Section 1.2.3), the number of studies that address the spatio-temporal dynamics of soil properties using ML methods is still limited. Our findings agree with the review by Grunwald (2009) who characterised studies covering the years 2007 and 2008. A big proportion of studies that deal with the temporal dynamics of soil properties are related to soil-water interactions, as shown in the Topic 3 of our topic detection analysis (Fig. 1.9).

We found three main approaches to deal with the temporal variation of soil:

Temporal extrapolation: The studies generate models for a specific time-step including one or more predictors that vary with time to then apply that fitted model to another time-step (e.g. Grinand *et al.*, (2017)).

Subtraction: The studies model two or more time-steps independently followed by a change analysis. For instance, Schillaci *et al.*, (2017b) and Zhang *et al.*, (2018b) subtracted the maps of the modelled properties from two different years to compute the change in SOC concentration and pH, respectively.

Dynamics: Studies that model the actual dynamics of a soil property based on some mechanistic or semi-mechanistic method. Stumpf *et al.*, (2018) created yearly land use covers for 8,500 km^2 in Switzerland using a combination of Landsat 5-7-8 and field land use observations in order to model the SOC dynamics based on the conversion regimes from their land use sequence patterns (Watson *et al.*, 2014).

Despite the availability of ML algorithms that have the capacity to capture 4D structures (e.g. Convolutional Recurrent Neural Networks), we did not find studies using ML to continuously model space and time simultaneously. We think the main reason is that soil observations are usually sparse in space-time (Grunwald, 2016) and

that it is not possible to fulfil the dataset size requirements of such models. That is the reason why we mostly find studies that use a mechanistic or semi-mechanistic approach.

1.2.6 Uncertainty assessment

Uncertainty assessment is an important requirement for any model, especially if the predictions are going to be used to guide decision-making. In this review, 24% of the studies, among most topics, present uncertainty assessment or mention the importance of considering it (Fig. 1.12).

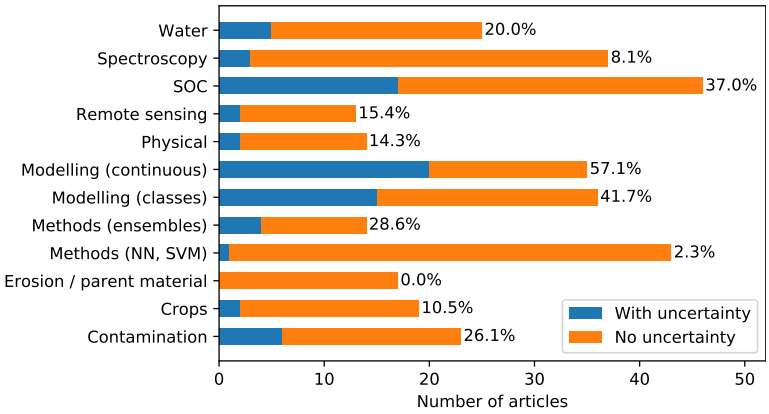


Fig. 1.12: Number of articles per topic that report or mention "uncertainty". Percentage represents proportion of total articles per topic.

In this review, a major contributor to the promotion of uncertainty assessment in soil modelling is the GlobalSoilMap project (Arrouays *et al.*, 2014) which, through specification developed by the DSM scientific community, recommends an uncertainty assessment of all their products. This is evident from Fig. 1.12, where topics related to DSM show a relatively high proportion of articles mentioning or reporting uncertainty. In the GlobalSoilMap specifications, the proposed uncertainty assessment method is the use of bootstraping when training the model (Stine, 1985), effectively making predictions with many models trained with subsets of the original data, to then estimate the 90% prediction interval (e.g. Castro-Franco *et al.*, (2017) and Ma *et al.*, (2017)). Another approach is the use of quantile regression (Koenker and Bassett Jr, 1978) to estimate the complete conditional distribution of the prediction. This method has

been recently applied in some DSM studies (Vaysse and Lagacherie, 2017; Sirsat *et al.*, 2018; Cao *et al.*, 2019). Less common approaches are the use of the fuzzy k-means with extragrades (Tranter *et al.*, 2010) algorithm, which defines areas within the covariate space, with different levels of uncertainty, were a new observation (to be predicted) can be placed; and the use of Bayesian optimisation approaches (Snoek *et al.*, 2015; Gal and Ghahramani, 2016).

1.3 General Discussion

1.3.1 Interpretability

Based on our findings, it is possible to state that, in general, ML methods have shown superior performance over more traditional methods in terms of predictive power. We now address the last questions from the aims of this paper — does an advanced model provide new insights that improve our knowledge and understanding of soils?

In order for a human to understand the decisions made by the model, the model has to be interpretable. The motivations for interpretability are varied, including trust, causality, transferability, informativeness and fairness (in ethical terms) (Lipton, 2016). In our review, researchers usually associate advanced ML models with low interpretability. For instance, Brungard *et al.*, (2015) assigned multiple models to different groups according to their complexity, with NN and SVM categorised as difficult to interpret compared to MLR or CART. Beguin *et al.*, (2017) also mention the lower interpretability of ML models compared with an explicit geostatistical model. Because the measurement of interpretability is usually not well defined, there are also contradictory opinions. For instance, RF is mostly considered in the category of low interpretability (Brungard *et al.*, 2015; Were *et al.*, 2015; Taghizadeh-Mehrjardi *et al.*, 2016; Deng *et al.*, 2018) but its use is also sometimes justified due to its ease of interpretability via the use of variables of importance (Jeong *et al.*, 2017).

It is important to clearly define the goal of a modelling exercise. If we want to obtain the model with the greatest accuracy in order to solve a specific problem, maybe interpretability should not be an important factor. If we consider a) that nature is a complex combination of nonlinear phenomena, and b) the limited capacity of humans to understand non-linear relationships (Doherty and Balzer, 1988), by requiring our

model complete transparency, we are limiting its capability. However, it is important to corroborate that the model is a valid generalisation of the studied phenomenon. If our goal is to obtain new insights, it is important to consider that interpretability goes hand in hand with prior knowledge and biases, and that we could be optimising an algorithm to present misleading but plausible explanations (Lipton, 2016).

How can we increase interpretability?

A common conclusion reported by authors of the reviewed papers is that the selection of the most informative or relevant predictors before training the model can increase interpretability (Xiong *et al.*, 2014; Prasad *et al.*, 2018; Wang *et al.*, 2018a; Keskin *et al.*, 2019), although some authors do not recommend selection of predictors based on the researchers' knowledge since it could lead to biased and suboptimal model performance (Brungard *et al.*, 2015; Keskin *et al.*, 2019). This discordance leads to a large range in the number of the predictors used, with some extreme cases using more than 200 (Xiong *et al.*, 2014; Keskin *et al.*, 2019).

NN are some of the best performing models but, given the complexity of their operation, they are usually labelled as "black-box" models. In consequence, many authors have focused on trying to provide frameworks to interpret the knowledge extracted by these models. For instance, Bau *et al.*, (2017) dissected a CNN to understand how different layers work and which features they favour by visualising their (neurons) activation map. Rauber *et al.*, (2017) used the activation maps projected into a 2D space in order to visualise and identify confusion zones, outliers, and clusters in the internal representations learned by the model.

In soil sciences, one of the reported methods to interpret ML models is to assess the importance of the variables used, usually derived from the number of times they have been used in the rules generated by tree-like models (Henderson *et al.*, 2005; Martin *et al.*, 2014; Schillaci *et al.*, 2017a; Khanal *et al.*, 2018). Another method to assess the relative influence of predictors in tree-like models is to estimate the average reduction of the error at each split of the tree, for all the predictors (Friedman, 2001). Another alternative, in the context of soil mapping, is to map the rules generated by the model to identify their spatial context or to map where important predictors were used (Bui *et al.*, 2006). For CNNs, by feeding simulated data to a trained model, Ng *et al.*,

(2019) explored the most important wavelengths used when predicting multiple soil properties from soil spectral data using a sensitivity analysis. The logic behind their analysis is that modifying unimportant wavelengths should not affect the prediction. By plotting the variance for the predictions by wavelengths it is possible to unveil the most important areas of the spectrum (Fig. 1.13).

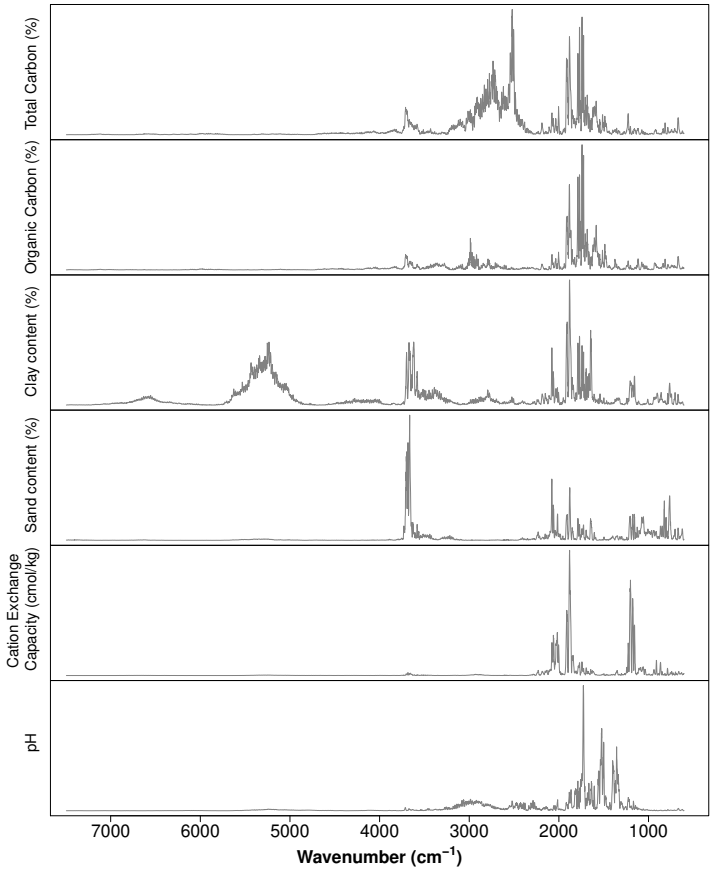


Fig. 1.13: Sensitivity analysis of CNN model prediction as a function of wavelength. Average variance of predictions by wavelength. This analysis allows to explore the most important wavelengths in a CNN model. Adapted from Ng *et al.*, (2019).

Interpretability is an important concept that should be revisited, since it is not absolute nor static, hence a specific model cannot be simply labelled as interpretable or not. Linear models can quickly become unintelligible as we add more variables (Lou *et al.*, 2012), and methods to better understand complex models such as NN are constantly being developed (Bau *et al.*, 2017; Montavon *et al.*, 2018; Zhang *et al.*, 2018a).

1.3.2 New good practices

Thanks to the effort of some groups to rescue soil legacy data (Arrouays *et al.*, 2017), and cheaper and faster methods to analyse soil samples, there is more soil data available than ever before. This data availability not only allows us to use new ML algorithms, which usually require more observations, but also opens the door to new ways to train those models. An important part of model development is validation. Literature traditionally recommends that an independent, unseen (by the model) dataset should be used as validation (Kohavi, 1995). In practice, the data is usually partitioned into training and validation datasets. A more stable solution is the use of k-cross-validation where the dataset is partitioned into k groups, where $k-1$ groups are used for training and 1 group for validation, repeating the training k times, each with a different validation group. When data availability is a limitation, researchers resort to techniques such as n-cross-validation or "leave-one-out" validation to make the most of the available data (Stevens *et al.*, 2008; Pasini, 2015).

A new generation of models based on NNs have been introduced in the later years, which have revolutionised many fields. Deep learning (DL) models, consisting of multiple hidden layers of neurons, have many parameters (from hundreds to millions) which need to be fitted in the training process. This is the reason why they usually need access to large sample sizes. A second characteristic of these models is that they have a considerable number of hyper-parameters. Hyper-parameters are parameters that are not learned from the data during the training phase and include things like number of iterations during the training, learning rate, layer parameters, number of layers, etc. A common practice when training DL models is to split the original dataset into 3 sub datasets: training, validation and test. The training dataset is used to learn the parameters, the validation dataset to compare models fitted with different hyper-parameters in order to find the optimal combination, and the test dataset as the independent, unseen data.

In soil sciences, ML algorithms are usually trained using the traditional train/validation split or cross-validation (Keskin *et al.*, 2019; Liang *et al.*, 2019), or even no validation (Feng *et al.*, 2019), except for some studies based on DL or with engineering background (e.g. Reale *et al.*, (2018)), including some of our publications on the use of DL for DSM (Padarian *et al.*, 2019c) or soil spectroscopy (Padarian *et*

al., 2019b; Padarian *et al.*, 2019a), which use a train/validation/test split. Considering the increasing size of datasets, we think soil scientist should transition towards the implementation of some DL practices such as dataset split and hyper-parameter optimisation (Bergstra and Bengio, 2012; Snoek *et al.*, 2012), not only for NNs but for any algorithm that has hyper-parameters. Some potential candidates are random forest, Cubist, classification and regression trees, and support vector machines. Most of the implementations of these algorithms have sensible default hyper-parameters but some studies report an important impact of them in their results (Mutanga *et al.*, 2012; Lu *et al.*, 2018). For general hyper-parameter tuning strategies, we refer the reader to Bergstra and Bengio (2012) for simple strategies such as grid or random search. For an in depth report of hyper-parameter tuning and its effects in the context of random forest, we refer the reader to Probst *et al.*, (2019).

1.3.3 Commercial ML applications

This work explores the use of ML in soil sciences by exploring the current scientific literature, but use of ML extends beyond research and companies are very welcoming to this technology, specially in applications such as computer vision, speech recognition, natural language processing, and robot control (Jordan and Mitchell, 2015). It is not hard to imagine a commercial application of approaches such as soil properties prediction using vis-NIR spectroscopy, either in the laboratory or the field. While in research there are some transparency requirements, including describing the methods and data used, companies are usually very secretive about their methods since they are a trade secret that gives them a competitive advantage. Considering that lack of transparency, how can we be sure that the predictions of their models are good? There is not a unique answer but it should include at least some uncertainty assessment (as discussed in Section 1.2.6) and information about the range of soils used during training.

In terms of the reporting soil types coverage, different approaches can be applied. A simple, perhaps over-confident method can be reporting the geographical extent from where the soil samples used during training were collected (e.g. Tomasella *et al.*, (2000) and Børgesen and Schaap (2005)), or a broad soil classification based in the soil characteristics such as “sandy soils” (e.g Schaap and Bouten (1996) and Shaw *et al.*,

(2000)). A better approach, based on the covariate space of the samples used during training is fuzzy k-means with extragrades, which has the benefit of describing both, coverage and uncertainty levels.

Even if uncertainty levels and coverage are reported, another factor to consider is how much we should trust in companies and their reports. Specially for applications involving public funding, but generally as a consumer protection measure, this type of products should be certifiable, in the same way many soil laboratories are. A usual approach is the use of reference materials (Dybczyński *et al.*, 1979; Pueyo *et al.*, 2001; Ahmed *et al.*, 2017), which should be consistent with the model coverage reported. The properties measured in the reference materials should fall within the prediction interval produced by the model, with a confidence defined for each application.

1.4 Conclusions and recommendations

Aided by a topic modelling approach, we were able to review the status of ML in soil sciences. We observed a general increase in the adoption of ML methods in time, and that its use is mostly concentrated in developed countries. This gap is probably due to the link between science, technology and development. We believe that proper inter-institutional collaboration plans should be put in place in order to close this gap.

By using topic modelling, we identified twelve categories of studies where ML is commonly used, namely remote sensing, soil organic carbon, water, contamination, methods (ensembles), erosion and parent material, methods (NN, SVM), spectroscopy, modelling (classes), crops, physical, modelling (continuous). The final topic model successfully captured relationships between topics such as modelling of continuous and categorical soil properties, and water, given that all these topics share a spatial component.

We also found that advanced ML methods usually perform better than simpler approaches thanks to their capability to capture non-linear relationships. However, it is important to note that more advanced methods usually require more data and that some precautions should be taken in order to avoid obtaining misleading results. Considering parsimony is always advised, hence if only a small, simple dataset is available, we recommend using a simple model. This also applies to the number of predictors. In consequence, according to many authors of the reviewed articles, it

is better to use meaningful predictors instead of relying on the model capabilities to "select the best variables" in order to improve interpretability.

Interpretability is an important aspect to consider when applying advanced ML methods in order to improve our knowledge and understanding of soil. Simpler methods (e.g. linear models) have been used for a long time and the way of interpreting them is well defined. More advanced methods (e.g. neural networks) are usually considered as "black box" models, but that is just a reflection of the current research state and not because it is impossible to interpret them. During our review, we found studies that proposed some solutions to improve their interpretability and we foresee that a large number of studies will focus on this topic.

1.5 References

Ahmad, S., Kalra, A., and Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources* 33 (1): 69–80.

Ahmed, O., Habbani, F. I., Mustafa, A., Mohamed, E., Salih, A., and Seedig, F. (2017). Quality assessment statistic evaluation of X-ray fluorescence via NIST and IAEA standard reference materials. *World Journal of Nuclear Science and Technology* 7 (02): 121.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., *et al.*, (2014). GlobalSoilMap: Toward a fine-resolution global grid of soil properties. In: *Advances in agronomy*. Vol. 125. Elsevier: pp. 93–134.

Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., *et al.*, (2017). Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ* 14: 1–19.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: pp. 6541–6549.

Beguin, J., Fuglstad, G.-A., Mansuy, N., and Paré, D. (2017). Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. *Geoderma* 306: 195–205.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13 (Feb): 281–305.

Björk, B.-C. (2017). Open access to scientific articles: a review of benefits and challenges. *Internal and emergency medicine* (2): 247–253.

Blanco, C. M. G., Gomez, V. M. B., Crespo, P., and Ließ, M. (2018). Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest. *Geoderma* 316: 100–114.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* 55 (4): 77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research* 3 (Jan): 993–1022.

Bondi, G., Creamer, R., Ferrari, A., Fenton, O., and Wall, D. (2018). Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation. *Geoderma* 318: 137–147.

Børgesen, C. D. and Schaap, M. G. (2005). Point and parameter pedotransfer functions for water retention predictions for Danish soils. *Geoderma* 127 (1-2): 154–167.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., and Edwards Jr, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma* 239: 68–83.

Bui, E. N., Henderson, B. L., and Viergever, K. (2006). Knowledge discovery from models of soil properties developed through data mining. *Ecological Modelling* 191 (3-4): 431–446.

Butler, B. M., O'Rourke, S. M., and Hillier, S. (2018). Using rule-based regression models to predict and interpret soil properties from X-ray powder diffraction data. *Geoderma* 329: 43–53.

Camera, C., Zomeni, Z., Noller, J. S., Zissimos, A. M., Christoforou, I. C., and Bruggeman, A. (2017). A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. *Geoderma* 285: 35–49.

Cao, B., Domke, G. M., Russell, M. B., and Walters, B. F. (2019). Spatial modeling of litter and soil carbon stocks on forest land in the conterminous United States. *Science of the Total Environment* 654: 94–106.

Castro-Franco, M., Domenech, M. B., Borda, M. R., Costa, J., *et al.*, (2017). Spatial dataset of topsoil texture for the southern Argentine Pampas. *Geoderma regional*.

Catlett, J. (1991). Mega induction: A test flight. In: *Machine Learning Proceedings 1991*. Elsevier: pp. 596–599.

Caubet, M., Dobarco, M. R., Arrouays, D., Minasny, B., and Saby, N. P. (2019). Merging country, continental and global predictions of soil texture: Lessons from ensemble modelling in France. *Geoderma* 337: 99–110.

Chlingaryan, A., Sukkarieh, S., and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture* 151: 61–69.

Coopersmith, E. J., Minsker, B. S., Wenzel, C. E., and Gilmore, B. J. (2014). Machine learning assessments of soil drying for agricultural planning. *Computers and electronics in agriculture* 104: 93–104.

Cortes, C., Jackel, L. D., Solla, S. A., Vapnik, V., and Denker, J. S. (1994). Learning curves: Asymptotic values and rate of convergence. In: *Advances in Neural Information Processing Systems*: pp. 327–334.

Costa, J. G., Reigosa, M., Matías, J., and Covelo, E. (2017). Soil Cd, Cr, Cu, Ni, Pb and Zn sorption and retention models using SVM: variable selection and competitive model. *Science of the Total Environment* 593: 508–522.

Dai, F., Zhou, Q., Lv, Z., Wang, X., and Liu, G. (2014). Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. *Ecological Indicators* 45: 184–194.

Deng, X., Chen, X., Ma, W., Ren, Z., Zhang, M., Grieneisen, M. L., Long, W., Ni, Z., Zhan, Y., and Lv, X. (2018). Baseline map of organic carbon stock in farmland topsoil in East China. *Agriculture, ecosystems & environment* 254: 213–223.

Dharumarajan, S, Hegde, R., and Singh, S. (2017). Spatial prediction of major soil properties using Random Forest techniques-A case study in semi-arid tropics of South India. *Geoderma Regional* 10: 154–162.

Dobarco, M. R., Cousin, I., Le Bas, C., and Martin, M. P. (2019). Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty. *Geoderma* 336: 81–95.

Doherty, M. E. and Balzer, W. K. (1988). Cognitive feedback. In: *Advances in psychology*. Vol. 54. Elsevier: pp. 163–197.

Dybczyński, R., Tugsavul, A, and Suschny, O. (1979). Soil-5, a new IAEA certified reference material for trace element determinations. *Geostandards Newsletter* 3 (1): 61–87.

Fajardo, M., McBratney, A., and Whelan, B. (2016). Fuzzy clustering of Vis–NIR spectra for the objective recognition of soil morphological horizons in soil profiles. *Geoderma* 263: 244–253.

Farfani, H. A., Behnamfar, F., and Fathollahi, A. (2015). Dynamic analysis of soil-structure interaction using the neural networks and the support vector machines. *Expert Systems with Applications* 42 (22): 8971–8981.

Feng, Y., Cui, N., Hao, W., Gao, L., and Gong, D. (2019). Estimation of soil temperature from meteorological data using different machine learning models. *Geoderma* 338: 67–77.

Flynn, T., Rozanov, A., de Clercq, W., Warr, B., and Clarke, C. (2019). Semi-automatic disaggregation of a national resource inventory into a farm-scale soil depth class map. *Geoderma* 337: 1136–1145.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*: 1189–1232.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *International conference on machine learning*: pp. 1050–1059.

Gao, M., Li, H.-Y., Liu, D., Tang, J., Chen, X., Chen, X., Blöschl, G., and Leung, L. R. (2018). Identifying the dominant controls on macropore flow velocity in soils: A meta-analysis. *Journal of Hydrology* 567: 590–604.

Geissen, V, Kampichler, C, López-de Llergo-Juárez, J., and Galindo-Acántara, A (2007). Superficial and subterranean soil erosion in Tabasco, tropical Mexico: development of a decision tree modeling approach. *Geoderma* 139 (3-4): 277–287.

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G., and Fernandes Filho, E. I. (2019). Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma* 340: 337–350.

Greifeneder, F., Khamala, E., Sendabo, D., Wagner, W., Zebisch, M., Farah, H., and Notarnicola, C. (2018). Detection of soil moisture anomalies based on Sentinel-1. *Physics and Chemistry of the Earth, Parts A/B/C*.

Grinand, C., Le Maire, G., Vieilledent, G., Razakamanarivo, H, Razafimbelo, T., and Bernoux, M. (2017). Estimating temporal changes in soil carbon stocks at ecoregional scale in Madagascar using remote-sensing. *International journal of applied earth observation and geoinformation* 54: 1–14.

Grunwald, S (2009). Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma* 152 (3-4): 195–207.

Grunwald, S. (2016). What do we really know about the space–time continuum of soil-landscapes? In: *Environmental Soil-Landscape Modeling*. CRC Press: pp. 16–49.

Grunwald, S., Vasques, G. M., and Rivero, R. G. (2015). Fusion of soil and remote sensing data to model soil properties. In: *Advances in Agronomy*. Vol. 131. Elsevier: pp. 1–109.

Han, J., Mao, K., Xu, T., Guo, J., Zuo, Z., and Gao, C. (2018). A soil moisture estimation framework based on the cart algorithm and its application in china. *Journal of hydrology* 563: 65–75.

Hanna, A. M., Ural, D., and Saygili, G. (2007). Neural network model for liquefaction potential in soil deposits using Turkey and Taiwan earthquake data. *Soil Dynamics and Earthquake Engineering* 27 (6): 521–540.

Heggemann, T., Welp, G., Amelung, W., Angst, G., Franz, S. O., Koszinski, S., Schmidt, K., and Pätzold, S. (2017). Proximal gamma-ray spectrometry for site-independent in situ prediction of soil texture on ten heterogeneous fields in Germany using support vector machines. *Soil and Tillage Research* 168: 99–109.

Henderson, B. L., Bui, E. N., Moran, C. J., and Simon, D. (2005). Australia-wide predictions of soil properties using decision trees. *Geoderma* 124 (3): 383–398.

Hutson, M. (2018). *Boycott highlights AI's publishing rebellion.*

Ivushkin, K., Bartholomeus, H., Bregt, A. K., Pulatov, A., Bui, E. N., and Wilford, J. (2018). Soil salinity assessment through satellite thermography for different irrigated and rainfed crops. *International journal of applied earth observation and geoinformation* 68: 230–237.

Jeong, G., Oeverdieck, H., Park, S. J., Huwe, B., and Ließ, M. (2017). Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. *Catena* 154: 73–84.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science* 349 (6245): 255–260.

Karandish, F. and Šimůnek, J. (2016). A field-modeling study for assessing temporal variations of soil-water-crop interactions under water-saving irrigation strategies. *Agricultural water management* 178: 291–303.

Keskin, H., Grunwald, S., and Harris, W. G. (2019). Digital mapping of soil carbon fractions with machine learning. *Geoderma* 339: 40–58.

Khadim, F. K., Su, H., Xu, L., and Tian, J. (2019). Soil salinity mapping in Everglades National Park using remote sensing techniques and vegetation salt tolerance. *Physics and Chemistry of the Earth, Parts A/B/C.*

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., and Shearer, S. (2018). Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and electronics in agriculture* 153: 213–225.

Kheir, R. B., Chorowicz, J., Abdallah, C., and Dhont, D. (2008). Soil and bedrock distribution estimated from gully form and frequency: A GIS-based decision-tree model for Lebanon. *Geomorphology* 93 (3-4): 482–492.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*: 33–50.

Koestel, J. and Jorda, H. (2014). What determines the strength of preferential transport in undisturbed soil under steady-state flow? *Geoderma* 217: 144–160.

Kohavi, R. *et al.*, (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*. Vol. 14. 2. Montreal, Canada: pp. 1137–1145.

Kovačević, M., Bajat, B., and Gajić, B. (2010). Soil type classification and estimation of soil properties using support vector machines. *Geoderma* 154 (3-4): 340–347.

Lacoste, M., Lemercier, B., and Walter, C. (2011). Regional mapping of soil parent material by machine learning based on point data. *Geomorphology* 133 (1-2): 90–99.

Leenaars, J. G., Claessens, L., Heuvelink, G. B., Hengl, T., González, M. R., van Bussel, L. G., Guilpart, N., Yang, H., and Cassman, K. G. (2018). Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa. *Geoderma* 324: 18–36.

Liang, Z., Chen, S., Yang, Y., Zhao, R., Shi, Z., and Rossel, R. A. V. (2019). National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. *Geoderma* 335: 47–56.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37 (1): 145–151.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Liu, S., Yang, Y., Shen, H., Hu, H., Zhao, X., Li, H., Liu, T., and Fang, J. (2018). No significant changes in topsoil carbon in the grasslands of northern China between the 1980s and 2000s. *Science of the total environment* 624: 1478–1487.

Lou, Y., Caruana, R., and Gehrke, J. (2012). Intelligible models for classification and regression. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM: pp. 150–158.

Lu, W., Lu, D., Wang, G., Wu, J., Huang, J., and Li, G. (2018). Examining soil organic carbon distribution and dynamic change in a hickory plantation region with Landsat and ancillary data. *Catena* 165: 576–589.

Ma, Y., Minasny, B., and Wu, C. (2017). Mapping key soil properties to support agricultural production in Eastern China. *Geoderma Regional* 10: 144–153.

Ma, Y., Minasny, B, Malone, B., and McBratney, A. (2019). Pedology and digital soil mapping (DSM). *European Journal of Soil Science*. . (Under review).

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., and Beaudoin, A. (2014). Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the k-nearest neighbor method. *Geoderma* 235: 59–73.

Märker, M., Pelacani, S., and Schröder, B. (2011). A functional entity approach to predict soil erosion processes in a small Plio-Pleistocene Mediterranean catchment in Northern Chianti, Italy. *Geomorphology* 125 (4): 530–540.

Martin, M., Orton, T., Lacarce, E, Meersmans, J, Saby, N., Paroissien, J., Jolivet, C, Boulonne, L, and Arrouays, D (2014). Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. *Geoderma* 223: 97–107.

Martinez, G., Weltz, M., Pierson, F. B., Spaeth, K. E., and Pachepsky, Y. (2017). Scale effects on runoff and soil erosion in rangelands: Observations and estimations with predictors of different availability. *Catena* 151: 161–173.

Massawe, B. H., Subburayalu, S. K., Kaaya, A. K., Winowiecki, L., and Slater, B. K. (2018). Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. *Geoderma* 311: 143–148.

Matthew and Honnibal, M. I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. . `https://github.com/explosion/spaCy/`.

McBratney, A., de Gruijter, J., and Bryce, A. (2019). Pedometrics timeline. *Geoderma* 338: 568–575.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Minasny, B and Flantis, D (2018). *"Helicopter research": who benefits from international studies in Indonesia?* `https://theconversation.com/helicopter-research-who-benefits-from-international-studies-in-indonesia-102165`. Accessed: 29/04/2019.

Mjolsness, E. and DeCoste, D. (2001). Machine learning for science: state of the art and future prospects. *science* 293 (5537): 2051–2055.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73: 1–15.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., and Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering* 152: 104–116.

Mutanga, O., Adam, E., and Cho, M. A. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation* 18: 399–406.

Naderi-Boldaji, M., Tekeste, M. Z., Nordstorm, R. A., Barnard, D. J., and Birrel, S. J. (2019). A mechanical-dielectric-high frequency acoustic sensor fusion for soil physical characterization. *Computers and Electronics in Agriculture* 156: 10–23.

Ng, W., McBratney, A., Minasny, B., Padarian, J., Monatzerolghaem, M., Ferguson, R., and Bailey, S. (2019). Deep learning for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*. . (Under review).

Oh, Y.-Y., Yun, S.-T., Yu, S., Kim, H.-J., and Jun, S.-C. (2019). A novel wavelet-based approach to characterize dynamic environmental factors controlling short-term soil surface CO2 flux: Application to a controlled CO2 release test site (EIT) in South Korea. *Geoderma* 337: 76–90.

Padarian, J, Minasny, B, and McBratney, A. (2019a). Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma* 340: 279–288.

Padarian, J, Minasny, B, and McBratney, A. (2019b). Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional* 16: e00198.

Padarian, J., Minasny, B., and McBratney, A. B. (2019c). Using deep learning for digital soil mapping. *Soil* 5 (1): 79–89.

Pasini, A. (2015). Artificial neural networks for small dataset analysis. *Journal of thoracic disease* 7 (5): 953.

Perlich, C., Provost, F., and Simonoff, J. S. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research* 4 (Jun): 211–255.

Poggio, L., Gimona, A., Spezia, L., and Brewer, M. J. (2016). Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA. *Geoderma* 277: 69–82.

Prasad, R., Deo, R. C., Li, Y., and Maraseni, T. (2018). Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. *Soil and Tillage Research* 181: 63–81.

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9 (3): e1301.

Pueyo, M, Rauret, G, Bacon, J., Gomez, A, Muntau, H, Quevauviller, P., and López-Sánchez, J. (2001). A new organic-rich soil reference material certified for its EDTA-and acetic acid-extractable contents of Cd, Cr, Cu, Ni, Pb and Zn, following collaboratively tested and harmonised procedures. *Journal of Environmental Monitoring* 3 (2): 238–242.

Rauber, P. E., Fadel, S. G., Falcao, A. X., and Telea, A. C. (2017). Visualizing the hidden activity of artificial neural networks. *IEEE transactions on visualization and computer graphics* 23 (1): 101–110.

Reale, C., Gavin, K., Librić, L., and Jurić-Kaćunić, D. (2018). Automatic classification of fine-grained soils using CPT measurements and Artificial Neural Networks. *Advanced Engineering Informatics* 36: 207–215.

Reeves, M. K., Perdue, M., Munk, L. A., and Hagedorn, B. (2018). Predicting risk of trace element pollution from municipal roads using site-specific soil samples and remotely sensed data. *Science of the Total Environment* 630: 578–586.

Rial, M, Cortizas, A. M., Taboada, T, and Rodríguez-Lado, L (2017). Soil organic carbon stocks in Santa Cruz Island, Galapagos, under different climate change scenarios. *Catena* 156: 74–81.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In: *Proceedings of the eighth ACM international conference on Web search and data mining*. ACM: pp. 399–408.

Rossiter, D. G. (2018). Past, present & future of information technology in pedometrics. *Geoderma* 324: 131–137.

Rudin, C. and Wagstaff, K. L. (2014). *Machine learning for science and society*.

Sagasti, F. R. (1973). Underdevelopment, science and technology: the point of view of the underdeveloped countries. *Science Studies* 3 (1): 47–59.

Schaap, M. G. and Bouten, W. (1996). Modeling water retention curves of sandy soils using neural networks. *Water Resources Research* 32 (10): 3033–3040.

Schillaci, C., Lombardo, L., Saia, S., Fantappiè, M., Märker, M., and Acutis, M. (2017a). Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. *Geoderma* 286: 35–45.

Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappie, M., Märker, M., and Saia, S. (2017b). Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. *Science of the total environment* 601: 821–832.

Shavlik, J. W., Mooney, R. J., and Towell, G. G. (1991). Symbolic and neural learning algorithms: An experimental comparison. *Machine learning* 6 (2): 111–143.

Shaw, J., West, L., Radcliffe, D., and Bosch, D. (2000). Preferential flow and pedotransfer functions for transport properties in sandy Kandiudults. *Soil Science Society of America Journal* 64 (2): 670–678.

Sirsat, M., Cernadas, E, Fernández-Delgado, M, and Barro, S (2018). Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods. *Computers and Electronics in Agriculture* 154: 120–133.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In: *Advances in neural information processing systems*: pp. 2951–2959.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. (2015). Scalable bayesian optimization using deep neural networks. In: *International conference on machine learning*: pp. 2171–2180.

Somarathna, P., Minasny, B., and Malone, B. P. (2017). More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. *Soil Science Society of America Journal*.

Song, X.-D., Yang, F., Ju, B., Li, D.-C., Zhao, Y.-G., Yang, J.-L., and Zhang, G.-L. (2018). The influence of the conversion of grassland to cropland on changes in soil organic carbon and total nitrogen stocks in the Songnen Plain of Northeast China. *Catena* 171: 588–601.

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual review of information science and technology* 41 (1): 643–681.

Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., and Ben-Dor, E. (2008). Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma* 144 (1-2): 395–404.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012). Exploring topic coherence over many models and many topics. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics: pp. 952–961.

Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association* 80 (392): 1026–1031.

Stumpf, F., Keller, A., Schmidt, K., Mayr, A., Gubler, A., and Schaepman, M. (2018). Spatio-temporal land use dynamics and soil organic carbon in Swiss agroecosystems. *Agriculture, ecosystems & environment* 258: 129–142.

Subburayalu, S., Jenhani, I, and Slater, B. (2014). Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. *Geoderma* 213: 334–345.

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., and Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. *Journal of the American Society for Information Science and Technology* 62 (1): 185–204.

Taghizadeh-Mehrjardi, R, Nabiollahi, K, Minasny, B, and Triantafilis, J (2015). Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. *Geoderma* 253: 67–77.

Taghizadeh-Mehrjardi, R, Nabiollahi, K, and Kerry, R (2016). Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. *Geoderma* 266: 98–110.

Tomasella, J., Hodnett, M. G., and Rossato, L. (2000). Pedotransfer Functions for the Estimation of Soil Water Retention in Brazilian Soils. *Soil Science Society of America Journal* 64: 327.

Tranter, G, Minasny, B., and McBratney, A. (2010). Estimating Pedotransfer Function Prediction Limits Using Fuzzy k-Means with Extragrades. *Soil Sci. Soc. Am. J.* 74 (6): 1967–1975.

Tziachris, P., Aschonitis, V., Chatzistathis, T., and Papadopoulou, M. (2019). Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. *Catena* 174: 206–216.

Vaysse, K. and Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* 291: 55–64.

Vincent, S., Lemercier, B., Berthier, L., and Walter, C. (2018). Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. *Geoderma* 311: 130–142.

Viscarra-Rossel, R. and Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 158 (1-2): 46–54.

Řehůřek, R. and Sojka, P. (May 2010). Software Framework for Topic Modelling with Large Corpora. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. `http://is.muni.cz/publication/884893/en`. Valletta, Malta: ELRA: pp. 45–50.

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li Liu, D., Simpson, M., McGowen, I., and Sides, T. (2018a). Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. *Ecological indicators* 88: 425–438.

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., and Li Liu, D. (2018b). High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. *Science of The Total Environment* 630: 367–378.

Ware, M. and Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing.

Warner, D. L., Guevara, M., Inamdar, S., and Vargas, R. (2019). Upscaling soil-atmosphere CO2 and CH4 fluxes across a topographically complex forested landscape. *Agricultural and forest meteorology* 264: 80–91.

Watson, S. J., Luck, G. W., Spooner, P. G., and Watson, D. M. (2014). Land-use change: incorporating the frequency, sequence, time span, and magnitude of changes into ecological research. *Frontiers in Ecology and the Environment* 12 (4): 241–249.

Were, K., Bui, D. T., Dick, Ø. B., and Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators* 52: 394–403.

Wu, G., Kechavarzi, C., Li, X., Wu, S., Pollard, S. J., Sui, H., and Coulon, F. (2013). Machine learning models for predicting PAHs bioavailability in compost amended soils. *Chemical engineering journal* 223: 747–754.

Wu, Q., Zhang, C., Hong, Q., and Chen, L. (2014). Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science* 40 (5): 611–620.

Xie, X.-L. and Li, A.-B. (2018). Identification of soil profile classes using depth-weighted visible–near-infrared spectral reflectance. *Geoderma* 325: 90–101.

Xing, L., Li, L., Gong, J., Ren, C., Liu, J., and Chen, H. (2018). Daily soil temperatures predictions for various climates in United States using data-driven model. *Energy* 160: 430–440.

Xiong, X., Grunwald, S., Myers, D. B., Kim, J., Harris, W. G., and Comerford, N. B. (2014). Holistic environmental soil-landscape modeling of soil organic carbon. *Environmental Modelling & Software* 57: 202–215.

Xu, Y., Smith, S. E., Grunwald, S., Abd-Elrahman, A., and Wani, S. P. (2017). Incorporation of satellite remote sensing pan-sharpened imagery into digital soil prediction and mapping models to characterize soil property variability in small agricultural fields. *ISPRS journal of photogrammetry and remote sensing* 123: 1–19.

Zeynoddin, M., Bonakdari, H., Ebtehaj, I., Esmaeilbeiki, F., Gharabaghi, B., and Haghi, D. Z. (2019). A reliable linear stochastic daily soil temperature forecast model. *Soil and Tillage Research* 189: 73–87.

Zhang, C., Mishra, D. R., and Pennings, S. C. (2019). Mapping salt marsh soil properties using imaging spectroscopy. *ISPRS Journal of Photogrammetry and Remote Sensing* 148: 221–234.

Zhang, Q., Nian Wu, Y., and Zhu, S.-C. (2018a). Interpretable convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: pp. 8827–8836.

Zhang, Y., Sui, B., Shen, H., and Wang, Z. (2018b). Estimating temporal changes in soil pH in the black soil region of Northeast China using remote sensing. *Computers and Electronics in Agriculture* 154: 204–212.

Zhou, D., Ji, X., Zha, H., and Giles, C. L. (2006). Topic evolution and social interactions: how authors effect research. In: *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM: pp. 248–257.

Appendix A – Articles count by journal

Table 1.A.1: List of journals by publisher and number of articles that marched the term 'soil "machine learning"' in a full-text search.

Journal	Articles
Geoderma	113
Science of The Total Environment	29
CATENA	18
Geoderma Regional	13
Computers and Electronics in Agriculture	12
Journal of Hydrology	11
Ecological Indicators	8
Remote Sensing of Environment	7
International Journal of Applied Earth Observation and Geoinformation	6
Agriculture, Ecosystems & Environment	5
Soil and Tillage Research	5
Journal of Terramechanics	5
Soil Biology and Biochemistry	5
Agricultural and Forest Meteorology	4
Computers & Geosciences	3
Chemometrics and Intelligent Laboratory Systems	3
Agricultural Water Management	3
Construction and Building Materials	3
ISPRS Journal of Photogrammetry and Remote Sensing	3
Forest Ecology and Management	3
Chemosphere	3
Environmental Modelling & Software	3
Environmental Pollution	3
Advanced Engineering Informatics	3
Geomorphology	3
Computers and Geotechnics	3

Continued on next page

Table 1.A.1: List of journals by publisher and number of articles that marched the term 'soil "machine learning"' in a full-text search.

Journal	Articles
Advances in Water Resources	3
Journal of Environmental Management	2
Soil Dynamics and Earthquake Engineering	2
Applied Soft Computing	2
Ecological Engineering	2
Journal of Geochemical Exploration	2
Sensors and Actuators A: Physical	2
Physics and Chemistry of the Earth, Parts A/B/C	2
Journal of Photochemistry and Photobiology B: Biology	1
Analytica Chimica Acta	1
Applied Geography	1
Applied Ocean Research	1
Chemical Geology	1
Information Processing in Agriculture	1
Geoscience Frontiers	1
Tunnelling and Underground Space Technology	1
Expert Systems with Applications	1
Ecological Modelling	1
Pedobiologia	1
Applied Radiation and Isotopes	1
Spectrochimica Acta Part B: Atomic Spectroscopy	1
iScience	1
Applied Geochemistry	1
Journal of Rock Mechanics and Geotechnical Engineering	1
Measurement	1
Environmental Technology & Innovation	1
Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy	1
Advances in Agronomy	1

Continued on next page

Table 1.A.1: List of journals by publisher and number of articles that marched the term 'soil "machine learning"' in a full-text search.

Journal	Articles
Environmental Research	1
Advances in Space Research	1
Journal of Hazardous Materials	1
Energy	1
Sustainable Computing: Informatics and Systems	1
Chemical Engineering Journal	1
Biosystems Engineering	1
Reliability Engineering & System Safety	1

Chapter 2

Using deep learning for Digital Soil Mapping

Summary

Digital soil mapping has been widely used as a cost-effective method for generating soil maps. However, current DSM data representation rarely incorporates contextual information of the landscape. DSM models are usually calibrated using point observations intersected with spatially corresponding covariates. Here, we demonstrate the use of the convolutional neural network model that incorporates contextual information surrounding an observation to significantly improve the prediction accuracy over conventional DSM models. We describe a convolutional neural network (CNN) model that takes inputs as images of covariates and explores spatial contextual information by finding non-linear local spatial relationships of neighbouring pixels. Unique features of the proposed model include: input represented as 3D stack of raster data, data augmentation to reduce overfitting, and simultaneously predicting multiple outputs. Using a soil mapping example in Chile, the CNN model was trained to simultaneously predict soil organic carbon at multiples depths across the country. The results showed that, in this study, the CNN model reduced the error by 30% compared with conventional techniques that only used point information of covariates. In the example of country-wide mapping at 100 m resolution, the neighbourhood size from 3 to 9 pixels is more effective than at a point location and larger neighbourhood sizes.

In addition, the CNN model produces less prediction uncertainty and it is able to predict soil carbon at deeper soil layers more accurately. Because the CNN model takes covariates represented as images, it offers a simple and effective framework for future DSM models.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. there are no other authors of the publication according to these criteria;

4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and

5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for this publication is:

Padarian, J., Minasny, B. and McBratney, A.B., 2019. Using deep learning for digital soil mapping. SOIL, 5(1), pp.79-89.

Contributors	Statement of contribution
José Padarian *Signature:* José Padarian *Date:* April 2, 2020	Conceptualisation Data analysis Writing
Budiman Minasny	Writing
Alex McBratney	Writing

2.1 Introduction

Digital soil mapping (DSM) has now been widely used globally for mapping soil classes and properties (Arrouays *et al.*, 2014). In particular, DSM has been used to map soil carbon efficiently around the world (e.g. Chen *et al.*, (2018)). DSM methodology has been adopted by FAO (FAO, 2018) so that digital soil maps can be produced reliably for sustainable land management. While DSM can now be said to be operational, there are still unresolved methodological issues regarding better representation of landscape pattern and soil processes. Some of the methodological research studies include the use of multiple remotely sensed images (Poggio and Gimona, 2017) or time series of images as covariates (Demattê *et al.*, 2018), testing novel regression and machine learning models (Angelini *et al.*, 2017; Somarathna *et al.*, 2017), and incorporation of spatial residuals of the regression model (Keskin and Grunwald, 2018; Angelini and Heuvelink, 2018).

The formalisation of the DSM methodology was done by the publication of McBratney *et al.*, (2003). Following the ideas of Dokuchaev (1883) and Jenny (1941), they described the *scorpan* model as the empirical quantitative relationship of a soil attribute and its spatially implicit forming factors. Such factors correspond to *a)* s: soil, other properties of the soil at a point; *b)* c: climate, climatic properties of the environment at a point; *c)* o: organisms, vegetation or fauna or human activity; *d)* r: topography, landscape attributes; *e)* p: parent material, lithology; *f)* a: age, the time factor; and *g)* n: space, spatial position. Explicitly, the *scorpan* model can be written as:

$$S_{(x,y)} = f(s_{(x,y)}, c_{(x,y)}, o_{(x,y)}, r_{(x,y)}, p_{(x,y)}, a_{(x,y)}, n_{(x,y)}) + e_{(x,y)} \tag{2.1}$$

where (x, y) corresponds to the coordinates of a soil observation, and e is the spatial residual.

The usual steps for deriving the *scorpan* spatial soil prediction functions include intersecting soil observations (point data) with the *scorpan* factors (raster images at a particular resolution), and calibrating a prediction function f. In effect, we are only looking at relationships between point observations and point representation of covariates. The *scorpan* factors have implicit spatial information, however the prediction function f does not explicitly take into account the spatial relationship.

Attempts have been made to incorporate more local information in the *scorpan* covariates, in particular topography. Approaches to include covariate information about the vicinity around the observations (x, y) have been devised. One approach is to derive topographic or terrain attributes (e.g. slope, curvature) at multiple scales by expanding the size of the window or neighbourhood size in the calculation (Miller *et al.*, 2015; Behrens *et al.*, 2010). Another approach includes multi-scale analysis using spatial filters such as wavelets on the covariate raster (Biswas and Si, 2011; Sun *et al.*, 2017). Thus, the raster represents larger spatial support. Studies indicated that, generally, covariates with larger support than its original resolution could enhance the prediction accuracy of the model (Mendonça-Santos *et al.*, 2006; Sun *et al.*, 2017).

DSM can be thought of as linking observable landscape structure and soil processes expressed as observed soil properties. To effectively link structure and processes, Deumlich *et al.*, (2010) suggested the use of analysis that spans over several spatial and temporal scales. Behrens *et al.*, (2018) proposed the contextual spatial modelling to account for the interactions of covariates across multiple scales and their influence on soil formation. The authors' approach (e.g. Behrens *et al.*, (2010) and Behrens *et al.*, (2014)) derived covariates based on the elevation at the local to the regional extent. Their approaches include ConMap (Behrens *et al.*, 2010) which is based on elevation differences from the centre pixel to each pixel in a sparse neighbourhood, and ConStat (Behrens *et al.*, 2014) which used statistical measures of elevation within growing sparse circular spatial neighbourhoods. These approaches produce a large number of predictors computed for each location, as shown in an example with 100 distance scales (e.g., from 20 m to 20 km) and 1000 predictors per grid cell. These hyper-covariates, solely based on elevation, are used as predictors in a random forest regression model.

Spatial filtering, multi-scale terrain calculation, and contextual mapping approaches require the pre-processing of each covariate independently. The useful scale for each covariate needs to be figured out via numerical experiments and most of the time the process relies on ad hoc decisions. Here, we take advantage of the success of deep learning models that are used for image recognition, as an effective tool in DSM to optimally search for local contextual information of covariates. This work aims to expand the classic DSM approach by including information about the vicinity around the observation (x, y), to fully leverage its spatial context. The aim is achieved

by devising a convolutional neural network (CNN) which can take multiple spatial contextual inputs.

2.2 Rationale

The theoretical background of DSM is based on the relationship between a soil attribute and soil forming factors. In practice, a single soil observation is usually described as a point p with coordinates (x, y) (Eq. 2.1), and the corresponding soil forming factors are represented by a vector of pixel values of multiple covariate rasters $(a_1, a_2, ..., a_n)$ at the same location, where n is the total number of covariate rasters.

Soils are highly dependent on their position in the landscape, and information at a particular pixel might not be sufficient to represent that complex relationship. Our method expands the classic DSM approach by replacing the covariates, usually represented as a vector, with a 3D array with shape (w, h, n), where w and h are the width and height in pixels of a window centred at point p (Fig. 2.1). Methods commonly used in DSM are not designed to adequately handle the data structure depicted in Fig. 2.1. The data representation is similar to the network model by Lee and Kwon (2017) which used hyperspectral images for classification purposes.

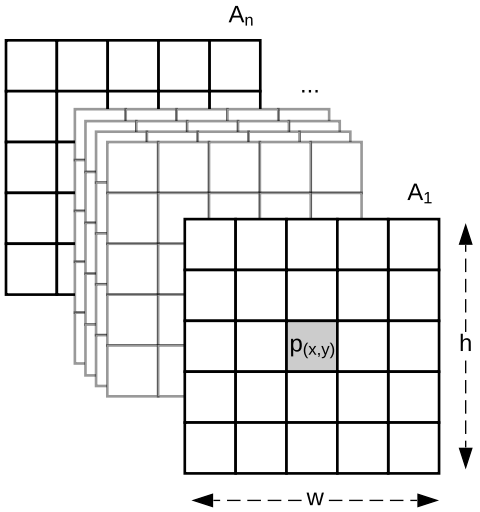


Fig. 2.1: Representation of the vicinity around a soil observation p, for n number of covariate rasters. w and h are the width and height in pixels, respectively.

As described in the introduction, while multi-scale or contextual mapping

approaches have been used in DSM, they still rely on a vector representation of covariates and rely on machine learning methods such as random forest to select important predictors. While deep learning methods have been used in DSM (e.g., Song *et al.*, (2016)), most studies still use a vector representation of covariates.

In the following sections, we introduce the use of convolutional neural networks (CNNs) to exploit spatial information of covariates that will perform a more effective DSM.

2.3 Deep learning

Deep learning is a machine learning method that is able to learn the representation of data through a series of processing layers. In agricultural and environmental mapping, it is mainly used in hyperspectral and multispectral image classification problems, e.g. land cover classification (Kamilaris and Prenafeta-Boldú, 2018). In this section we briefly introduce CNNs and some associated methods used during this work. For a more detailed and general description about CNNs we refer the reader to LeCun *et al.*, (1990) and Krizhevsky *et al.*, (2012).

2.3.1 CNN

CNNs are based on the concept of a layer of convolving windows which move along a data array in order to detect features (e.g. edges) of the data by using different filters (Fig. 2.2). When stacked together, convolutional layers are capable of extracting features of increasing complexity and abstraction (LeCun *et al.*, 1990). Since CNNs have the capacity to leverage the spatial structure of the data, they have been widely and effectively used in computer vision for image recognition or extraction (LeCun and Bengio, 1995).

A CNN has a number of three dimensional hidden layers, each layer learning to detect different features of the input images (LeCun *et al.*, 2015). In our case, each of the layers can perform one of two types of operations: convolution, or pooling. Convolution takes the input images through a set of convolutional filters (e.g., a 3x3 size filter), each of which detects and enhances certain features from the images. Units in a convolutional layer are organised in feature maps (here we used 3 x 3). Each unit of

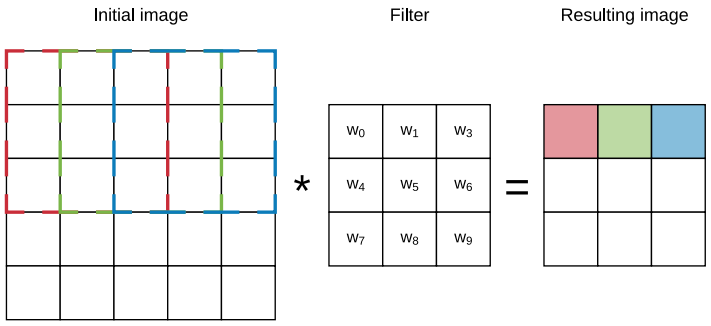


Fig. 2.2: Example of the first 3 steps of a convolution of a 3x3 filter over a 5x5 array (image). The resulting pixel values correspond to the sum of the element-wise multiplication of the initial pixels (dashed lines) and the filter.

the feature map is connected to local patches in the feature maps of the previous layer through a set of weights. The local weighted sum is then passed through a non-linear transfer function.

A pooling operation merges similar features by performing non-linear down-sampling. Here we used Max-Pooling layers which combine inputs from a small 2x2 window. Pooling also makes the features robust against noise. All the convolutional and pooling layers are finally "flattened" to the fully connected layer. In effect, the fully connected layer is a weighted sum of the previous layers.

To obtain optimal weights for the network, we train the network using a training data set. Weights were adjusted based on a gradient-based algorithm to minimise the error using an Adam optimiser (Kingma and Ba, 2014). We refer to a review by LeCun *et al.*, (2015) on the details of CNN.

2.3.2 Multi-task learning

CNNs have the capacity to predict multiple properties simultaneously. By doing so, a multi-task CNNs is capable of sharing learned representations between different targets and also can use the other targets as "clues" during the prediction process. In consequence, the error of the simultaneous prediction is generally lower compared with a single prediction for each target (Padarian *et al.*, 2019; Ramsundar *et al.*, 2015). An additional advantage of using a multi-task CNN is the reported reduction in the risk of overfitting (Ruder, 2017).

In DSM, where the combination of large extents, high resolution, and bootstrap routines, leads to running multiple model realisations on billions of pixels, combined with the fact that CNNs use a group of pixels around the soil observation instead of a single pixel, the time and computational resources required for training and inference is an important factor. Due to the simultaneous training and inference of multiple targets, a multi-task CNN presents the advantage of reducing both, training and inference time, compared with a single-task model.

2.4 Methods

2.4.1 Data

The data used in this work correspond to Chilean soil information. Since most observations are distributed on agricultural lands, we complemented that information with a second small data collection compiled from the literature and collaborators. We selected soil organic carbon (SOC) content (%) at depths 0–5, 5–15, 15–30, 30–60 and 60–100 cm as our target attributes. A total of 485 soil profiles were used after excluding soil profiles with total thickness greater than 100 cm (in order to assure that all the profiles have observations at all depth intervals). For more details about the data and depth standardisation we refer the reader to Padarian *et al.*, (2017).

As covariates, we used a) digital elevation model (HydroSHEDS, Lehner *et al.*, (2008)), which are provided at 3 arc-second resolution, in addition to its derived slope and topographic wetness index, calculated using SAGA (Conrad *et al.*, 2015); and b) long term mean annual temperature and total annual rainfall derived from information provided by WorldClim (Hijmans *et al.*, 2005), at 30 arc-second resolution. All data layers were standardised to a 100 m grid size.

2.4.2 Data augmentation

Deep learning techniques are described as "data-hungry" since they usually work better with large volumes of data. The direct effect of data augmentation is to generate new samples by modifying the original data without changing its meaning (Simard *et al.*, 2003). To achieve this, we rotated the 3D array shown in Fig. 2.1 by 90, 180, and 270

degrees, hence quadruplicating the number of observations. It is important to note that the central pixel preserves its initial position. The model trained on augmented data was compared with a model trained on the data without augmentation.

A secondary effect of data augmentation is regularisation, reducing the variance of the model and overfitting (Krizhevsky *et al.*, 2012). Data augmentation also induces rotation invariance (Vo and Hays, 2016) by generating alternative situations (rotated data) where the model response should be similar to the original data (e.g: a soil profile next to a gully is expected to be similar to a profile next to the opposite side of the gully, *ceteris paribus*).

2.4.3 Network architecture

The multi-task CNN used in this study (Fig. 2.3; Table 2.1) consists of an input layer pass through a series of convolutional and pooling layers with a ReLU activation function, which adds a non-linearity by passing the learned weights through the function $f(x) = max(0, x)$. The initial common/shared network has a function of extracting features shared between the five target depth ranges. Next, the common features are propagated through independent branches, one per depth range, of 3 fully-connected layers.

Table 2.1: Sequence of layers used to build the multi-task neural network

Layer type	Kernel size	Filters	Activation
†Convolutional	3x3	16	ReLU
†Max-Pooling	2x2	-	-
†Dropout (0.3)	-	-	-
†Convolutional	3x3	32	ReLU
‡Fully-connected	-	10	ReLU
‡Dropout (0.3)	-	-	-
‡Fully-connected	-	10	ReLU
‡Fully-connected	-	1	ReLU

†Shared layers; ‡for each property

The multiple connection between the layers generates a high number of parameters. In order to reduce the risk of overfitting, we introduce a dropout rate. In between the layers, 0.3 of the connections were randomly disconnected (Nitish *et al.*, 2014).

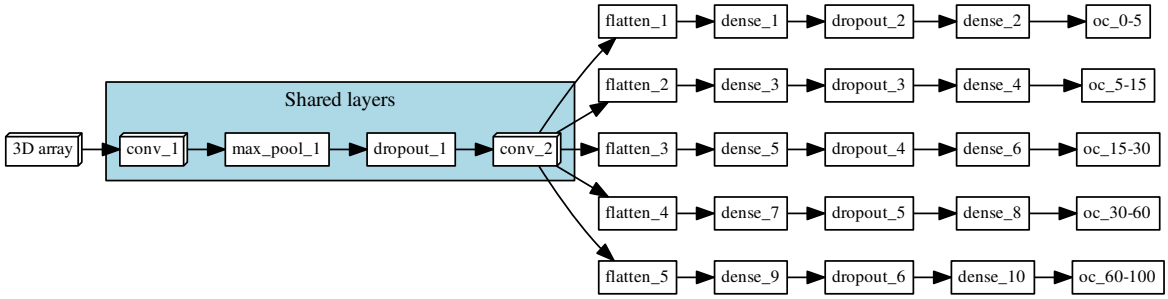


Fig. 2.3: Architecture of the multi-task network. "Shared layers" represent the layers shared by all the depth ranges. Each branch, one per depth range, first flattens the information to a 1D array, followed by a series of 2 fully-connected layer and a fully-connected layer of size=1, which corresponds to the final prediction.

We added this dropout operation in the shared layer and another dropout before the output.

2.4.4 Inputs

As explained in Section 2.2, our method uses a window around a soil observation which encloses a group of pixels instead of the single pixel that coincides with the observation. Most likely, the extent or size of that window will affect the model performance. To assess this effect, we compared the results of different models trained with a window size of 3, 5, 7, 9, 15, 21 and 29 pixels.

As the vicinity size increases, so does the number of parameters of the network (considering a fixed network architecture) and the risk of overfitting. To minimise overfitting, we modified the architecture of the network depending on the vicinity size (Table 2.2).

2.4.5 Training & Validation

First, 10% (n = 49) of the total dataset was randomly selected and used as a test set. The remaining 90% of the samples (n = 436) were augmented (see Section 2.4.2) obtaining a total of 1,744 samples. Following the data augmentation, we performed a bootstrapping routine (Efron and Tibshirani, 1993) with 100 repetitions, where the training set is obtained by sampling with replacement, generating a set of size

Table 2.2: List of modifications made to the base network architecture for specific input window sizes.

Window size	Changes
15x15	• Extra Max-Pooling(2x2) after last Convolutional layer
21x21	• Extra Max-Pooling(2x2) after last Convolutional layer • Extra Convolutional(3x3, 16 filters)
29x29	• Extra Max-Pooling(2x2) after last Convolutional layer • Extra Convolutional(3x3, 64 filters) • Droputs changed to 0.5

1,744. The samples which were not selected, about 1/3 (one-third), correspond to the out-of-bag validation set.

We compared our results with a previous study by Padarian *et al.*, (2017) where we used a Cubist regression tree model (Quinlan, 1992) to predict SOC at a national extent. Cubist has been used in many other DSM studies due to its interpretability and robustness. In that study, we used the same set of soil observations and covariates described in Section 2.4.1.

2.4.6 Uncertainty analysis

In this work (and in Padarian *et al.*, (2017)), the uncertainty is represented as the 90% prediction interval derived from the 100 bootstrap iterations. To estimate the upper and lower prediction interval limits, we used the formula:

$$PIL = \bar{x} \pm 1.645\sqrt{\sigma^2 + MSE} \tag{2.2}$$

where $\bar{x}$ and σ^2 are the mean and variance of the 100 iterations, per pixel, and MSE is the mean square error of the 100 fitted models.

2.4.7 Implementation

The CNN was implemented in Python (v3.6.2; Python Software Foundation, 2017) using Keras (v2.1.2; Chollet, 2015) and Tensorflow (v1.4.1; Abadi *et al.*, 2015) backend.

Computing was done using the University of Sydney's Artemis high performance computing facility.

2.5 Results and discussion

2.5.1 Data augmentation

Data augmentation was effective at reducing model error and variability (Fig. 2.4). It was possible to observe a decrease of the RMSE, by 10.56, 10.56, 11.25, 14.51, and 24.77% for 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. The results are in accordance with image classification studies which generally showed that data augmentation increased the accuracy of classification tasks (Perez and Wang, 2017). It is hypothesised that by increasing the amount of training data, we can reduce overfitting of CNN models.

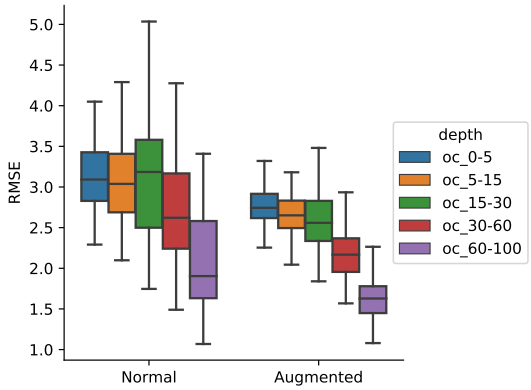


Fig. 2.4: Effect of using data augmentation as a pre-treatment on a 7x7 pixels array. The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR)

In this particular case, data augmentation partially adds rotational invariance, which makes sense when we don't expect the model to capture the effect of covariates such as aspect. If the model were to include a covariate such as aspect, this type of data augmentation (rotation) should not be applied. In terms of the data spatial autocorrelation, we need to consider that after augmenting the data we have 4 samples in the same locations with exactly the same SOC content, therefore assuming that

there is no variance when distance=0. That is theoretically true if we consider that the distance is exactly equal to 0. In practice, when calculating the semivariogram, the semivariance value of the first bin will be lower, but that does not significantly affect the final spatial autocorrelation of the process.

2.5.2 Vicinity size

To incorporate contextual information for DSM prediction, we represent the input as an image. The image is represented as an observation in the centre, with surrounding pixels in a square format. The size of the neighbourhood window (vicinity) has a significant effect on the prediction error (Fig. 2.5). There is no significant difference when using a vicinity size of 3, 5, 7 or 9 pixels, but sizes above 9 pixels showed an increase in the error. It is possible to observe a lower error in the test dataset, compared with the training and validation, due to the slight differences in the dataset distributions (Fig. 2.6). Since the SOC distribution is right-skewed, the random sampling used to generate the training dataset does not completely recreate the original distribution, excluding samples with very high SOC values. This should not significantly affect the conclusions given that the error for the samples with high SOC values is accounted for during the bootstrapping routine and reflected in the training and validation curves of Fig. 2.5. In this example, for a country-scale mapping of SOC at 100 m grid size, information from 150 to 450 m radius is useful. A similar influence distance was obtained by Jian-Bing *et al.*, (2006) and Sun *et al.*, (2003) whom reported a medium-scale spatial correlation range for SOC in China of around 300 m and 550 m, respectively; Rossi *et al.*, (2009) and the 190 m reported for a Coastal forest in Tanzania; and Don *et al.*, (2007) of around 200 m in two grassland sites in Germany. A similar spatial correlation range was reported for croplands in a review by Paterson *et al.*, (2018), where, based on 41 variograms, the authors estimated an average spatial correlation range of around 400 m.

As described in Section 2.4.3, we slightly modified the architecture of the network as input window size increased, in order to minimise the risk of overfitting and isolate the effect of the vicinity size. As we increase the vicinity size, we give the model a broader spatial context. Our results show that just a small amount of extra context provides enough information to improve the model predictions, and a larger amount

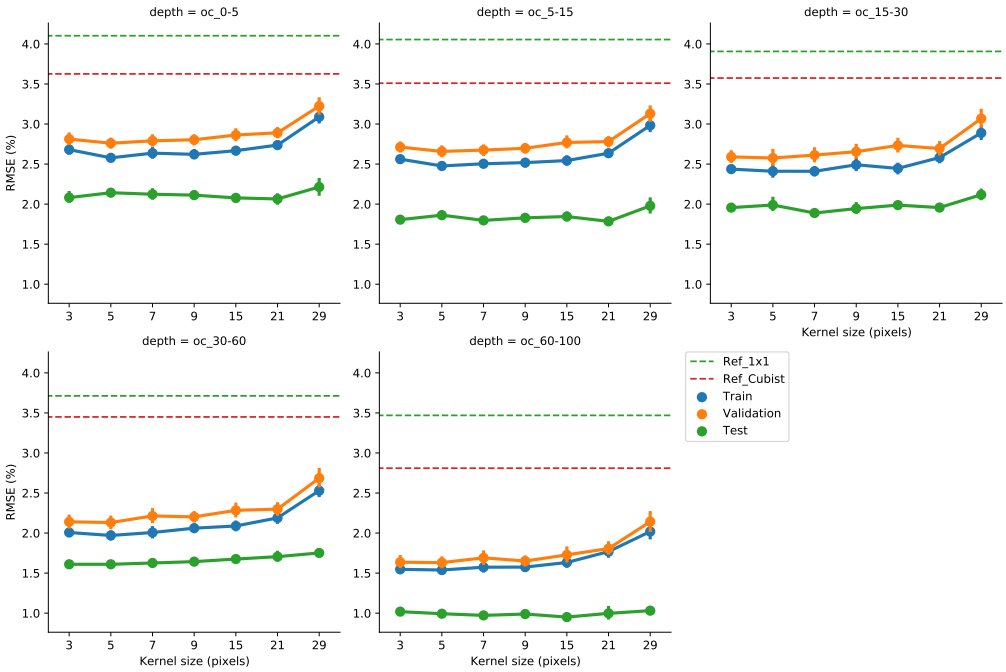


Fig. 2.5: Effect of vicinity size on prediction error, by depth range. Ref_1x1 corresponds to a fully connected neural network without any surrounding pixels. Ref_Cubist corresponds to the Cubist models used by Padarian *et al.*, (2017).

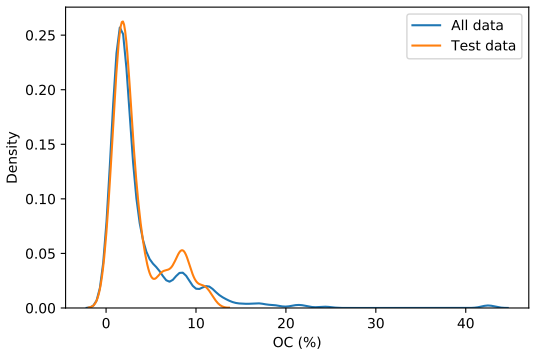


Fig. 2.6: Distribution of the original dataset and the test dataset. Density as a proportion of the total number of samples. Note that the random sampling excludes some observation with high SOC values.

of neighbouring information acts as noise, impairing the generalisation of the model. Since we used the relatively large resolution of 100 m, it is hard to tell specifically what is the minimum amount of context needed to improve SOC predictions. We believe that using higher resolutions (< 10m) could produce more insights about this matter.

Soil forming factors interact in complex ways and affect soil properties with different strength. At local scale, a broader context (i.e. larger vicinity size) does not necessarily provide extra information to the model, for instance when one of the factors is relatively homogeneous. The extra information could be even detrimental if the vicinity size is well beyond the area of influence of a factor, which is what probably happened when we increased the vicinity size above 9 pixels (radius≈450m). Representing this complexity in numerical terms would imply varying the size of the input array, such as each forming factor has a different vicinity size, most likely also varying depending on the spatial location of the soil observation (e.g. smaller vicinity for homogeneous areas, larger for heterogeneous areas). This is technically possible but considerably increases the complexity of the modelling.

2.5.3 Comparison with other methods

We compared our approach with the Cubist model used in our previous study (Padarian *et al.*, 2017), where we did not use any contextual information. We observed a significant decrease in the error (Fig. 2.5) by 23.0, 23.8, 26.9, 35.8, 39.8% for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. Most current DSM studies rely on punctual observations without contextual information and, given the improvements shown by our approach, we believe there is a big potential for CNNs to be used in operational DSM.

To compare our results with a method that uses contextual information, we ran a test using wavelet decomposition as per Mendonça-Santos *et al.*, (2006). In addition to the five covariates, we used their approximation coefficients from the first, second and third levels of a Haar decomposition (Chui, 2016; Haar, 1910). The results including wavelet decomposed variables were similar to ones obtained with the Cubist model. The CNN approach reduced the error by 24.8, 24.7, 28.5, 28.6, 23.5% for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. Mendonça-Santos *et al.*, (2006) reported an average improvement of 1% for the prediction of clay content. In our case the wavelet decomposition reduced the error of SOC content by 5.1%, in average, compared with Cubist but the reduction was only observed in depth (2.4, 1.2, 2.3, -10.1, -21.4% error change for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively), where SOC content is low, hence reducing its effectiveness in

applications such as carbon accounting. Our approach showed greater error reductions and through the whole profile.

2.5.4 Prediction of deeper soil layers

Our approach uses a multi-task CNN to predict multiple depths simultaneously in order to produce a synergistic effect. Compared with predicting each depth range in isolation by training a network with the same structure (Section 2.4.3) but with only one output, our approach reduced the error by 1.5, 6.7, 6.6, 8.9, 13.0% for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. In this case, the reduction was modest and we believe the effect can be greater when more soil observations are available.

In DSM, there are two main approaches to deal with the vertical variation of a soil attribute: 2.5D and 3D modelling. In the first one, an independent model is fitted for each depth range. The latter explicitly incorporates depth in order to obtain a single model for the whole profile. Interestingly, both approaches show a decrease in the variance explained by the model as the prediction depth increases. In a 3D mapping of SOC for a 125 km^2 region in the Netherlands, Kempen *et al.*, (2011) presented R^2 values of 0.75, 0.23 and 0.09 for the 0–30, 30–60, and 60–90cm depth ranges, respectively. In our previous study (Padarian *et al.*, 2017) the 2.5D mapping showed R^2 values of 0.39, 0.39, 0.27, 0.19, and 0.17 for the 0–5, 5–15, 15–30, 30–60 and 60–100 cm depth ranges, respectively. Similar studies show the same trend (Akpa *et al.*, 2016; Mulder *et al.*, 2016; Adhikari *et al.*, 2014), independent of the models used or the soil attribute predicted. This is expected as the information used as covariates usually represents surface conditions. Our multi-task network presented the opposite trend (Fig. 2.7), showing an increase of the explained variance as the prediction depth increases.

The prediction of the adjacent layers served as guidance, producing a synergistic effect. A soil attribute through a profile usually has a predictable behaviour (unless there are lithological discontinuities), which has been described by many authors in the form of depth functions (Kempen *et al.*, 2011; Nakane, 1976; Russell and Moore, 1968). A CNN is capable of generating an internal representation of the vertical distribution of the target attribute, which resembles the observed pattern (Fig. 2.8).

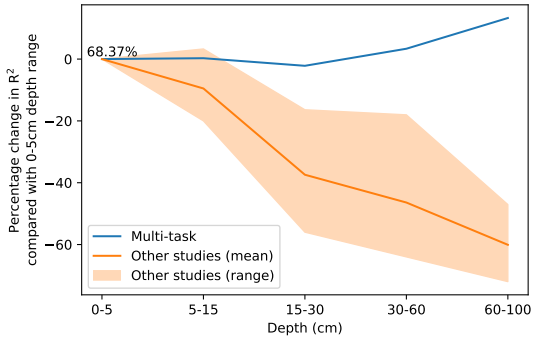


Fig. 2.7: Percentage change in model R^2 in function of depth. The Multi-task model corresponds to a CNN trained using a 7x7 pixels vicinity. Data for "Other studies" correspond to validation statistics from Padarian *et al.*, (2017), Akpa *et al.*, (2016), and Mulder *et al.*, (2016) and Adhikari *et al.*, (2014)

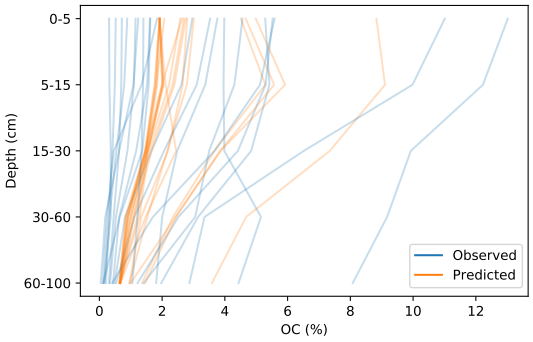


Fig. 2.8: Vertical SOC distribution for 20 randomly selected profiles. Predictions correspond to the multi-task CNN.

2.5.5 Visual evaluation of maps

Visually, the maps generated with the Cubist tree model and our multi-task CNN showed differences (Fig. 2.9). In an example for an area in southern Chile (around 72.57° S), the map generated with the Cubist model (Fig. 2.9a) shows more details related with the topography, but also presents some artefacts due to the sharp limits generated by the tree rules. On the other hand, the map generated with the multi-task CNN using a 7x7 window (Fig. 2.9b) shows a smoothing effect, an expected behaviour as a consequence of using neighbour pixels.

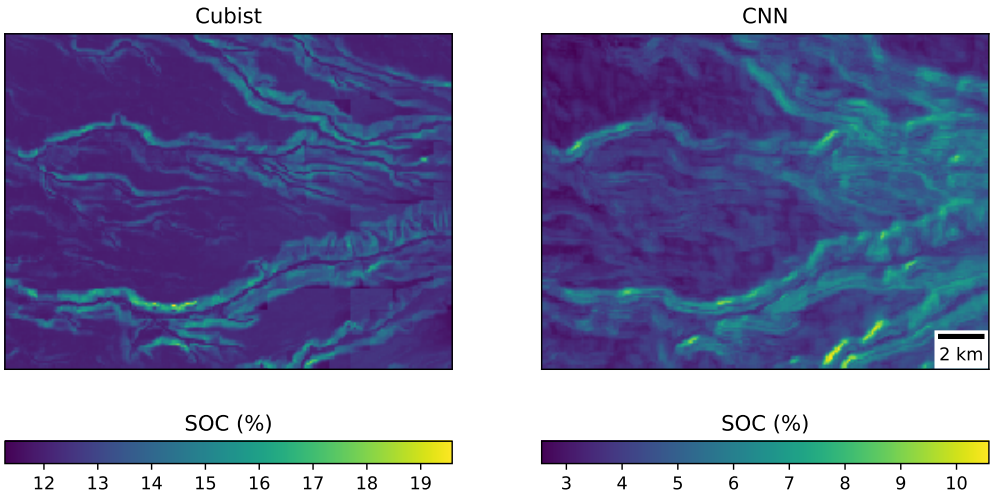


Fig. 2.9: Detailed view of (left panel) map generated by a Cubist model (Padarian *et al.*, 2017) and (right panel) model generated by our multi-task CNN showing the smoothing effect of the CNN. The maps correspond to the 0–5cm depth interval.

2.5.6 Uncertainty

A recommended DSM practice is to generate a map of a predicted attribute along with its associated uncertainty (Arrouays *et al.*, 2014). Our multi-task CNN significantly reduced the prediction interval width (PIW, Table 2.3) compared with the Cubist model. On average, we observed a reduction of 13.1 and 13.8% for the CNN model generated without and with data augmentation pre-treatment, respectively, for the first three depth intervals. Our multi-task CNN model showed a slightly lower prediction interval coverage, but all wider than the proposed 90% coverage.

In terms of the spatial patterns of the uncertainty (Fig 2.10), the greater reductions of the PIW were observed in elevated areas of the Andes, followed by the central valleys. A slight increase, in the order of 6-8%, was observed in the western coastal ranges. The reduction of the PIW in the Andes is most likely due to a more reserved extrapolation by the CNN models compared with Cubist. It is worth noting that the central valleys is where most of the agricultural lands are located and the uncertainty reduction observed in these areas could have important implications.

Table 2.3: Median prediction interval width (PIW, SOC %) and proportion of observations that fell within the 90% prediction interval (PICP) estimated at the test dataset locations. For the Cubist model, values were extracted from the final maps. For the CNN models, the values correspond to the mean of the 100 bootstrap iterations.

		Cubist	Not augmented	Augmented
0-5 cm	PICP	0.96	0.96	0.94
	PIW	7.96	7.20	7.25
5-15 cm	PICP	0.97	0.96	0.92
	PIW	7.69	6.15	6.06
15-30 cm	PICP	0.97	0.96	0.96
	PIW	7.16	6.47	6.35

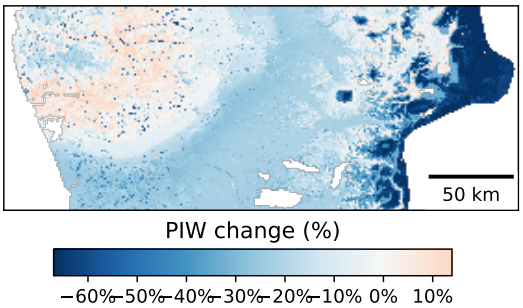


Fig. 2.10: Percentage change on the prediction interval width when using our CNN (with data augmentation) versus a Cubist model.

2.6 Conclusions

Incorporating contextual information into DSM models is an important aspect that deserves more attention. Since a soil surveyor will look at the surrounding landscape to make a prediction of soil type, DSM models should also incorporate information surrounding an observation. We demonstrated the use of a convolutional neural network as an efficient, effective, and accurate method to achieve this goal. In particular we introduce a deep learning model for DSM which has the following innovative features:

- The representation of input as an image, which takes into account information surrounding a point observation. CNN is able to recognise

contextual information, and extract multi-scale information automatically, which circumvents the need to pre-process the data in the form of spatial filtering or multi-scale analysis.

- The use of data augmentation as a general representation of soil in the landscape, which can reduce overfitting and improve model accuracy.
- The ability to predict different soil depths simultaneously in a model, and thus take into account the depth correlation of soil properties and attributes. In our example, prediction of soil properties at deeper depths, common problem in DSM studies, improved significantly.

Overall, in this study, we observed an error reduction of 30% compared with conventional techniques. The resulting prediction also has less uncertainty. Furthermore, the use of this data structure with CNN seems to eliminate artefacts generally found in DSM products due to incompatible scale of covariates and sharp discontinuities due to tree models.

A CNN can handle large numbers of covariates and has advantages over other machine learning algorithms used in DSM, such as random forests, and Cubist regression tree because its architecture is flexible, and explicitly takes spatial information of covariates around observations. While there have been attempts to include information surrounding an observation as covariates in a random forest model, those inputs still do not have spatial relationships. CNN does not require pre-processing such as wavelet transformation, rather such functions are built in the model. There are other features such as handling missing values via data imputation (Duan *et al.*, 2016) which can be readily added in the network.

The example presented in this paper is for a country-wide modelling at 100 m resolution, and we need to further test such an approach in regional to landscape mapping. The CNN model would be highly suitable for mapping soil class. In addition, the presented model can be used for other environmental mapping.

2.7 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org.

Adhikari, K., Hartemink, A. E., Minasny, B., Kheir, R. B., Greve, M. B., and Greve, M. H. (2014). Digital mapping of soil organic carbon contents and stocks in Denmark. *PloS one* 9 (8): e105519.

Akpa, S. I., Odeh, I. O., Bishop, T. F., Hartemink, A. E., and Amapu, I. Y. (2016). Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma* 271: 202–215.

Angelini, M. E. and Heuvelink, G. B. (2018). Including spatial correlation in structural equation modelling of soil properties. *Spatial Statistics* 25: 35–51.

Angelini, M., Heuvelink, G., and Kempen, B (2017). Multivariate mapping of soil with structural equation modelling. *European Journal of Soil Science* 68 (5): 575–591.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., *et al.*, (2014). GlobalSoilMap: Toward a fine-resolution global grid of soil properties. In: *Advances in agronomy*. Vol. 125. Elsevier: pp. 93–134.

Behrens, T, Schmidt, K, MacMillan, R., and Rossel, R. V. (2018). Multiscale contextual spatial modelling with the Gaussian scale space. *Geoderma* 310: 128–137.

Behrens, T., Schmidt, K., Zhu, A.-X., and Scholten, T. (2010). The ConMap approach for terrain-based digital soil mapping. *European Journal of Soil Science* 61 (1): 133–143.

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.-X., and Scholten, T. (2014). Hyper-scale digital soil mapping and soil formation analysis. *Geoderma* 213: 578–588.

Biswas, A. and Si, B. C. (2011). Revealing the controls of soil water storage at different scales in a hummocky landscape. *Soil Science Society of America Journal* 75 (4): 1295–1306.

Chen, S., Martin, M. P., Saby, N. P., Walter, C., Angers, D. A., and Arrouays, D. (2018). Fine resolution map of top-and subsoil carbon sequestration potential in France. *Science of The Total Environment* 630: 389–400.

Chollet, F. *et al.*, (2015). *Keras*. `https://github.com/fchollet/keras`.

Chui, C. K. (2016). *An introduction to wavelets*. Elsevier.

Conrad, O, Bechtel, B., Bock, M, Dietrich, H, Fischer, E, Gerlitz, L., Wehberg, J, Wichmann, V, and Böhner, J. (2015). System for automated geoscientific analyses (SAGA) v. 2.1.4. *Geoscientific Model Development* 8 (7): 1991–2007.

Demattê, J. A. M., Fongaro, C. T., Rizzo, R., and Safanelli, J. L. (2018). Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sensing of Environment* 212: 161–175.

Deumlich, D., Schmidt, R., and Sommer, M. (2010). A multiscale soil–landform relationship in the glacial-drift area based on digital terrain analysis and soil attributes. *Journal of Plant Nutrition and Soil Science* 173 (6): 843–851.

Dokuchaev, V. V. (1883). Russian Chernozem. Selected works of V.V. Dokuchaev. v. 1. *Israel Program for Scientific Translations*. . Jerusalem (translated in 1967).

Don, A., Schumacher, J., Scherer-Lorenzen, M., Scholten, T., and Schulze, E.-D. (2007). Spatial and vertical variation of soil carbon at two grassland sites—implications for measuring soil carbon stocks. *Geoderma* 141 (3-4): 272–282.

Duan, Y., Lv, Y., Liu, Y.-L., and Wang, F.-Y. (2016). An efficient realization of deep learning for traffic data imputation. *Transportation research part C: emerging technologies* 72: 168–181.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Vol. 57. New York: CRC press: p. 436.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen* 69 (3): 331–371.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* 25 (15): 1965–1978.

Jenny, H. (1941). Factors of soil formation: a system of quantitative pedology New York. *Macgraw Hill*.

Jian-Bing, W., Du-Ning, X., Xing-Yi, Z., Xiu-Zhen, L., and Xiao-Yu, L. (2006). Spatial variability of soil organic carbon in relation to environmental factors of a typical small watershed in the black soil region, northeast China. *Environmental monitoring and assessment* 121 (1-3): 597–613.

Kamilaris, A. and Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture* 147: 70–90.

Kempen, B., Brus, D., and Stoorvogel, J. (2011). Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. *Geoderma* 162 (1-2): 107–123.

Keskin, H and Grunwald, S (2018). Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma* 326: 22–41.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*: pp. 1097–1105.

LeCun, Y., Bengio, Y., *et al.*, (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361 (10): 1995.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In: *Advances in neural information processing systems*: pp. 396–404.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553): 436–444.

Lee, H. and Kwon, H. (2017). Going deeper with contextual CNN for hyperspectral image classification. *IEEE Transactions on Image Processing* 26 (10): 4843–4855.

Lehner, B., Verdin, K., and Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data. *EOS, Transactions American Geophysical Union* 89 (10): 93–94.

McBratney, A., Mendonça Santos, M. L., and Minasny, B (2003). On digital soil mapping. *Geoderma* 117 (1): 3–52.

Mendonça-Santos, M., McBratney, A., and Minasny, B (2006). Soil prediction with spatially decomposed environmental factors. *Developments in Soil Science* 31: 269–278.

Miller, B. A., Koszinski, S., Wehrhan, M., and Sommer, M. (2015). Impact of multi-scale predictor selection for modeling soil properties. *Geoderma* 239: 97–106.

Mulder, V. L., Lacoste, M., de Forges, A. R., and Arrouays, D. (2016). GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth. *Science of the total environment* 573: 1352–1369.

Nakane, K. (1976). An empirical formulation of the vertical distribution of carbon concentration in forest soils. *Japanese Journal of Ecology* 26 (3): 171–174.

Nitish, S., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15 (1): 1929–1958.

Padarian, J, Minasny, B, and McBratney, A. (2017). Chile and the Chilean soil grid: a contribution to GlobalSoilMap. *Geoderma Regional*: 17–28.

Padarian, J, Minasny, B, and McBratney, A. (2019). Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional* 16: e00198.

Paterson, S., McBratney, A. B., Minasny, B., and Pringle, M. J. (2018). Variograms of Soil Properties for Agricultural and Environmental Applications. In: *Pedometrics*. Springer: pp. 623–667.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Poggio, L. and Gimona, A. (2017). Assimilation of optical and radar remote sensing data in 3D mapping of soil properties over large areas. *Science of the Total Environment* 579: 1094–1110.

Python Software Foundation (2017). *Python Language Reference*. Python Software Foundation.

Quinlan, J. R. *et al.*, (1992). Learning with continuous classes. In: *5th Australian joint conference on artificial intelligence*. Vol. 92. Singapore: pp. 343–348.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*.

Rossi, J., Govaerts, A., De Vos, B., Verbist, B., Vervoort, A., Poesen, J., Muys, B., and Deckers, J. (2009). Spatial structures of soil organic carbon in tropical forests—a case study of Southeastern Tanzania. *Catena* 77 (1): 19–27.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Russell, J. and Moore, A. (1968). Comparison of different depth weightings in the numerical analysis of anisotropic soil profile data. *Int Soc Soil Sci Trans*.

Simard, P. Y., Steinkraus, D., Platt, J. C., *et al.*, (2003). Best practices for convolutional neural networks applied to visual document analysis. In: *ICDAR*. Vol. 3: pp. 958–962.

Somarathna, P., Minasny, B., and Malone, B. P. (2017). More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. *Soil Science Society of America Journal*.

Song, X., Zhang, G., Liu, F., Li, D., Zhao, Y., and Yang, J. (2016). Modeling spatio-temporal distribution of soil moisture by deep learning-based cellular automata model. *Journal of Arid Land* 8 (5): 734–748.

Sun, B., Zhou, S., and Zhao, Q. (2003). Evaluation of spatial and temporal changes of soil quality based on geostatistical analysis in the hill region of subtropical China. *Geoderma* 115 (1-2): 85–99.

Sun, X.-L., Wang, H.-L., Zhao, Y.-G., Zhang, C., and Zhang, G.-L. (2017). Digital soil mapping based on wavelet decomposed components of environmental covariates. *Geoderma* 303: 118–132.

Vo, N. N. and Hays, J. (2016). Localizing and orienting street views using overhead imagery. In: *European Conference on Computer Vision*. Springer: pp. 494–509.

Chapter 3

Online machine learning for collaborative biophysical modelling

Summary

Many initiatives try to integrate data from different parties to solve problems that could not be addressed by a sole participant. Despite the well-known benefits of collaboration, concerns of data privacy and confidentiality are still an obstacle that impedes progress in collaborative global research. This work tackles this issue using an online-learning algorithm to generate a single model where the data remains with each party and there is no need to integrate it to a single source. This approach is demonstrated in building a global soil organic carbon model based on databases of field observations held by 65 different countries. The model is trained by visiting each country, one at a time. Only knowledge and parameters of the model are transferred between countries. The results show that it is possible that the proposed approach yields a similar prediction accuracy compared with a model that is trained with all the data.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. there are no other authors of the publication according to these criteria;

4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and

5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for this publication is:

Padarian, J., Minasny, B. and McBratney, A.B., 2019. Online machine learning for collaborative biophysical modelling. Environmental Modelling & Software, 122, p.104548.

Contributors	Statement of contribution
José Padarian *Signature:* José Padarian *Date:* April 2, 2020	Conceptualisation Data analysis Writing
Budiman Minasny	Writing
Alex McBratney	Writing

3.1 Introduction

The rise of big data in various disciplines has presented new opportunities to address global-scale health and environmental issues by data integration, analysing data compiled from various countries or agencies. Examples include compiling a global gene database (DeSantis *et al.*, 2003), compiling a global soil database (Batjes, 2009), and analysing species-interaction datasets (Poelen *et al.*, 2014). In an ideal situation, data sharing that allows full access to datasets collected by different parties can enhance collaboration between researchers and accelerate finding solutions. Data sharing is even associated with increased citation rates (Piwowar *et al.*, 2007). Despite the positive nature of this collaborative spirit, data sharing comes with concerns about privacy and confidentiality (Fienberg *et al.*, 1985), and national sovereignty.

The data privacy issue is commonly attended by preventing data sharing. It is critical to develop techniques to enable the integration and sharing of data without losing privacy. Some privacy-preserving methods that have been proposed include data aggregation, adding noise to the original data (Wieland *et al.*, 2008), or data encryption (Kantarcioglu and Clifton, 2004). Others include data sanitising, or removing privacy-sensitive information from the database which could be time-consuming.

As an alternative to these sometimes convoluted solutions, we propose a novel use of an online learning technique via machine learning (ML) algorithms. We use the concept of transferring "knowledge" extracted from the data but not transferring the data itself, thus preserving data privacy. ML has been widely used in biophysical modelling with first appearances in the late 1980s - early 1990s (Yost *et al.*, 1988; Stockwell *et al.*, 1990). This paper explores the use of ML algorithms to generate a single model with data horizontally distributed between different parties (i.e., each party has different observations or rows). Each data holder party cannot see other parties' data, but yet the model can be built using all parties' data.

The proposed solution can aid researchers of different disciplines of earth and environmental sciences which require collaborative projects. The major challenge in such collaborative projects is data privacy. Bulatewicz *et al.*, (2013) proposed a general-purpose software component as a foundation for model linkages and integrated studies where data is "pulled" from other sources to feed a model (within a model chain). Data privacy issues can still be a concern in such an approach. The "pulling"

step can be replaced by our approach to obtain a distributed chain of models with data partitioned between different parties. Our approach could also be applied to modular approaches like SME3 (Spatial Modeling Environment 3; Maxwell (1999)), or any model integration framework (Dolk and Kottemann, 1993). More generally, our approach can be used in many other collaborative projects where difficulties when compiling data are common (Borgman, 2012) but rarely reported.

First, we present the theory of online learning algorithms and its optimisation strategy. Then we explore the use of stochastic gradient descent, a optimisation technique, as a way of optimising the model from distributed data. Finally, we demonstrate the application of online learning as a way of building a global soil carbon model where the model visits each country successively.

3.2 Online learners

The basis of ML is to optimise parameters of the model from the data. The most common way to find the optimal parameters for a given model is by minimising an objective function. ML has a variety of methods to optimise this process, including some modifications of the classical "steepest descent method" proposed by Cauchy (1847). Batch gradient descent (BGD) is a variation of this method which uses all the available data to find its way, always pointing in the right direction to a minimum objective function. The successive estimates w_t of an optimal parameter w are computed by

$$w_{t+1} = w_t - \eta_t \frac{1}{L} \sum_{i=1}^{L} \nabla_w Q\left(z_i, w_t\right) \tag{3.1}$$

where $\nabla_w Q$ is the gradient of the loss function Q, z an observation (x, y) composed of a set of covariates x and a response y, η_t the learning rate, and L the total number of observations. At each iteration of the BGD algorithm, the average of the gradients of the loss function needs to be calculated using the entire dataset.

A variation of BGD, stochastic gradient descent (SGD), approximates to the solution using a subsample of the training data (sub-batch) at a time, usually in a more erratic trajectory compared with BGD. In the case when the batch size is equal to one, this is done by computing the successive estimates w_t using

$$w_{t+1} = w_t - \eta_t \nabla_w Q(z_t, w_t) \tag{3.2}$$

By dropping the averaging operation in Eq. 3.1, the SGD algorithm can optimise a model without the need to access all the data at once. In theory, averaging all the small optimisation updates would yield a similar result to the one obtained by BGD. For a more complete description of both algorithms, we refer the reader to Bottou (1998).

Given this characteristic, SGD has been widely used in ML to solve large scale problems (Zhang, 2004; Bousquet and Bottou, 2008), where the amount of data is too big to be handled by the core memory of a computer system (Toledo, 1999), or in applications where it is necessary to process a stream of information like financial (stock) data (Lee *et al.*, 2010).

Considering a case where every sub-batch is visited once, the use of the SGD optimisation can be expanded to train a model where every sub-batch resides in different, independent locations. In this work, we focus on horizontally partitioned data, where each party contributes the same data attributes in different rows.

3.3 A case of mapping global soil carbon stock

We illustrate the use of the proposed online machine learning algorithm with a case study inspired by a recent publication by Stockmann *et al.*, (2015). The study aimed to create a global model to predict soil organic carbon (SOC), based on environmental covariates and a curated database of field observations from various countries. Soil is recognised as the largest terrestrial carbon store, and the knowledge of its distribution and stock allows implementation of global initiatives such as the 4 per 1000 initiative which was launched during the COP 21 meeting, with an aim to increase global soil carbon content to offset anthropogenic greenhouse gas emission (Minasny *et al.*, 2017).

The modelling was performed within a digital soil mapping framework where a soil property is a function of spatially referenced environmental covariates that represent soil-forming factors (McBratney *et al.*, 2003). Digital soil mapping researchers, in collaboration with different organisations, have done an exhaustive work "rescuing" and organising soil legacy data in many countries (Arrouays *et al.*, 2017). This data has been used to generate high resolution maps of predicted key soil properties along with

their prediction uncertainties. Much of the rescued soil legacy data has been shared via soil information systems (e.g. SISLAC (CIAT, 2013), SPADE/M2 (Hiederer, 2010)) or national databases but an important proportion has not, mainly due to legislative or license restrictions.

The dataset used in this work comprises 106,431 observations distributed among 65 countries (Fig. 3.1). Each of the data contributing countries have at least 100 observations. The aim is to develop a model that predicts the topsoil SOC content (% of OC at 0-5 cm depth) using global environmental covariates: elevation (Danielson and Gesch, 2011) and derived slope, long-term mean annual temperature and total annual precipitation (Hijmans *et al.*, 2005). All the covariates were resampled to 500m resolution.

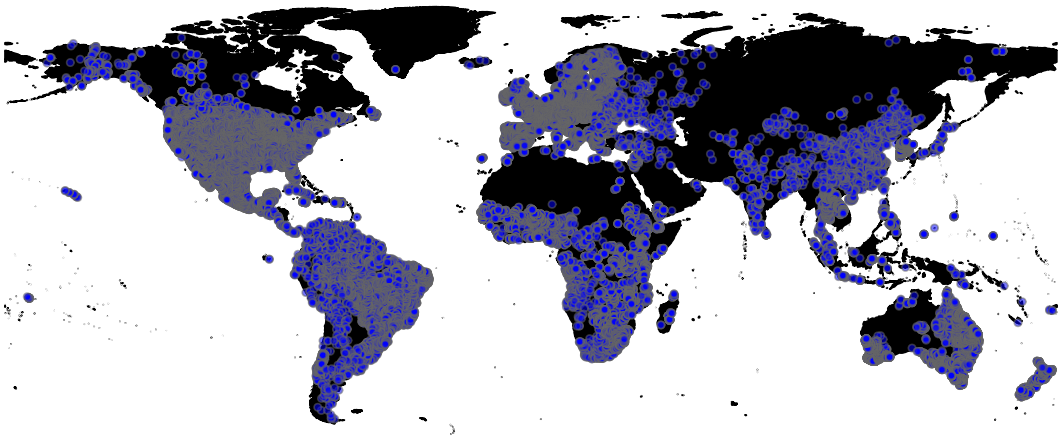


Fig. 3.1: Location of observations. Points are semi-transparent, hence intense blue areas have overlapping symbols.

To generate the model, we used the scikit-learn Python library (Pedregosa *et al.*, 2011). We selected a neural network regressor (MLPRegressor) as our model since neural networks have proven to be a flexible alternative (Sarle, 1994) yielding good results in many disciplines. The model architecture consisted of one 100-neurons hidden layer and a $ReLU$ activation function (Nair and Hinton, 2010). We used the SGD optimiser using a sub-batch size of 5, maximum 200 iterations, an initial learning rate η_0 of $5\text{x}10^{-4}$ (hyperparameters selected after a grid search considering the final model performance. See Section 3.4.4 for a discussion about an alternative approach), and an inverse scaling learning rate schedule, where the effective learning rate η_t at each

iteration t is given by:

$$\eta_t = \frac{\eta_0}{\sqrt{t}}. \tag{3.3}$$

We trained the MLPRegressor by feeding data to the model one country at a time. This procedure simulates the situation where, at each step, the partially trained model "travels" to a new location to learn from a new dataset. The whole process is described in Fig. 3.2.

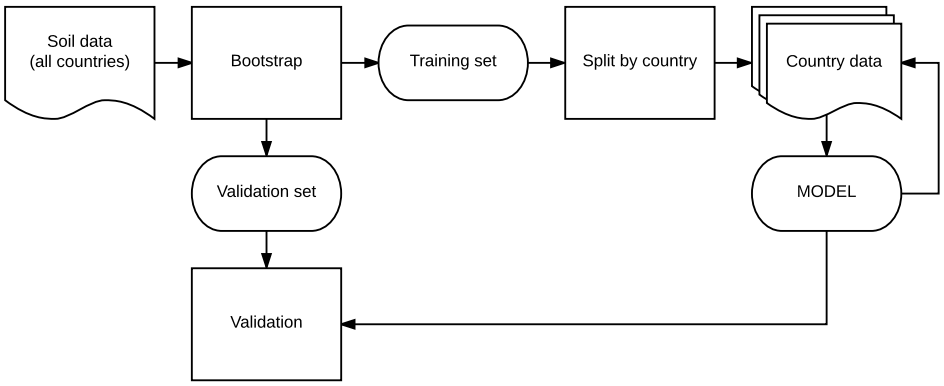


Fig. 3.2: Training and validation pipeline. Note that the model only has access to one country at a time during the training process.

To facilitate convergence, the predictors were scaled before training the model and, in this study, the covariates were scaled to the 0–1 range, using the minimum and maximum values of the first country added to the queue, as in

$$x_i' = \frac{x_i - \min(x_1)}{\max(x_1) - \min(x_1)} \qquad i = \{2, 3, ..., n\} \tag{3.4}$$

where x_1 corresponds to the values of the covariate x for the first country added to the queue, and x_i to the values of the covariate x for the subsequent countries. The reason to perform such scaling is that it is not possible to know *a-priori* the distribution of the subsequent datasets to be fed to the model. In practice, if the scaled values of the subsequent countries are outside the 0–1 range, is not a problem since the values will still be relatively small. To circumvent the effect of the order in which the data were fed to the model, we used a random permutation of the order of the countries. Before training the model, the SOC values were log-transformed and the model predictions

were back-transformed before proceeding with the analysis.

3.3.1 A comparison with an alternative approach

A possible alternative to our method is the use of model ensemble, specifically averaging the predictions of different learners, which is a common way of combining models (Breiman, 1996). We trained a model for each country separately. Following the same order as the online model training, we averaged the predictions of the first n models, where n is equal to the number of countries already processed. The prediction of the global model is generated by an ensemble of the models ($\hat{y}$):

$$\hat{y} = \sum_{i=1}^{n} y_i w_i \tag{3.5}$$

where y_i is the prediction of the model i, and w_i corresponds to the weight assigned to that model. In this work we used an equal weight averaging, where $w_i = \frac{1}{n}$.

In addition to the average, at each step, we calculated the total variance of the predictions following error propagation rules (Birge, 1939).

3.3.2 Response to redundant data

The training pipeline described above simulates a situation where the data came from different countries and where each country occupies a certain spatial domain, thus adding new information from each country will complete a global picture of the model. In an uncontrolled training condition, the data can come from anywhere. To simulate this condition and assess the sensitivity of the proposed approach to redundant information, we added five randomly selected subsets of already-seen data, repeated from one to six times, to the end of the training queue.

3.3.3 Evaluation

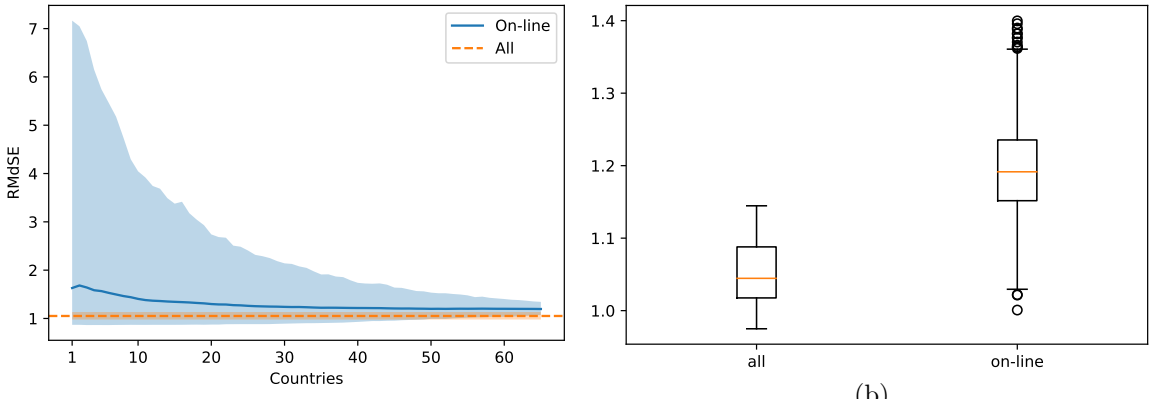
We performed 1,000 realisations of the model using the statistical bootstrap sampling with replacement (Efron and Tibshirani, 1993). At each of the incremental training steps, the model was evaluated against observations not included on the bootstrap sample by calculating the root of the median squared error (RMdSE). We decided

to use RMdSE instead of the root mean squared error since the SOC distribution is heavily right skewed and a measure based on the median squared error is a more robust alternative (Kempen *et al.*, 2012) to evaluate the overall effect of using online learning. To compare our proposed approach with the single global dataset approach, we trained the same algorithm using the complete dataset (from all of the countries) with 1,000 realisations in a bootstrap routine.

3.4 Results and discussion

3.4.1 Online versus complete dataset

The simulation results showed that the online model, which trained the model using data from one country at a time, decreases the error successively, and, at the end of the routine, reaches an error level that is slightly larger than the model trained with the complete dataset (Fig. 3.3). More notably, the variation of the model predictions, given by the bootstrap sampling and also the random order in which the countries were added, decreases considerably as we added more countries.



(a) (b)

Fig. 3.3: (a) Mean RMdSE (% SOC) for each consecutive training. Shaded areas correspond to the the difference between the 97.5 and 2.5 percentiles. Dashed line correspond to the RMdSE of the the reference model trained on all the data at once. (b) RMdSE (% SOC) after the last country is added to the model. The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR)

Both error and variation of the prediction depict the evolution of the online model. At the first stage, the model is specific and very sensitive to the initial data (the country that initiated the model) and, as the training evolves with more countries adding their data, the model becomes more general and robust. This condition is more evident when we observe the evolution of the prediction error of the first and last country added to the model (Fig. 3.4). The first stage of the model represents a model which is calibrated from a country's data, which performed poorly when it is extrapolated to the rest of the world. As data from more countries are added to the training queue, the model becomes more universal and, as a consequence, the prediction error of the first country added to the queue starts increasing (as it loses that country's specificity). The opposite effect was observed for the last country added to the model, which has a large error at the beginning of the model evolution, and decreases as more knowledge was added into the model.

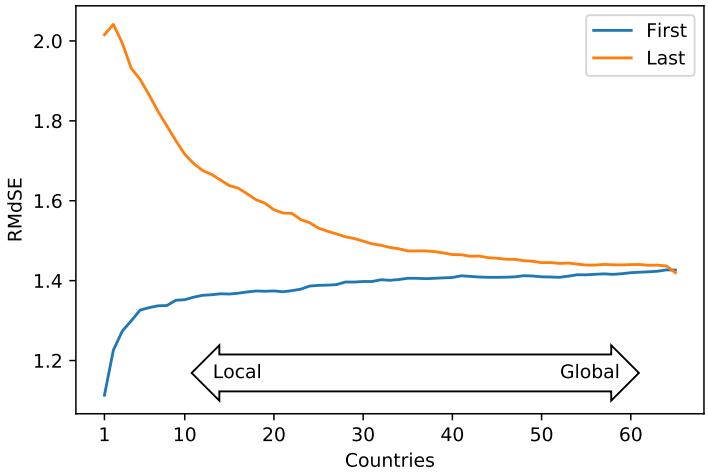


Fig. 3.4: RMdSE (% SOC) for the first and last country added to the queue, for each consecutive training. The value corresponds to the mean of the 1,000 iterations.

In terms of the spatial representation of the model (Fig. 3.5), the pattern is similar to the evolution of the model error as described in Fig. 3.4. Initial maps generated using knowledge extracted from few countries' data do not show the known pattern of global SOC distribution. The maps evolved as knowledge from different countries were incorporated into the model, and the spatial distribution of SOC converged to the established global pattern Stockmann *et al.*, (2015) as depicted in the final step of Fig. 3.5.

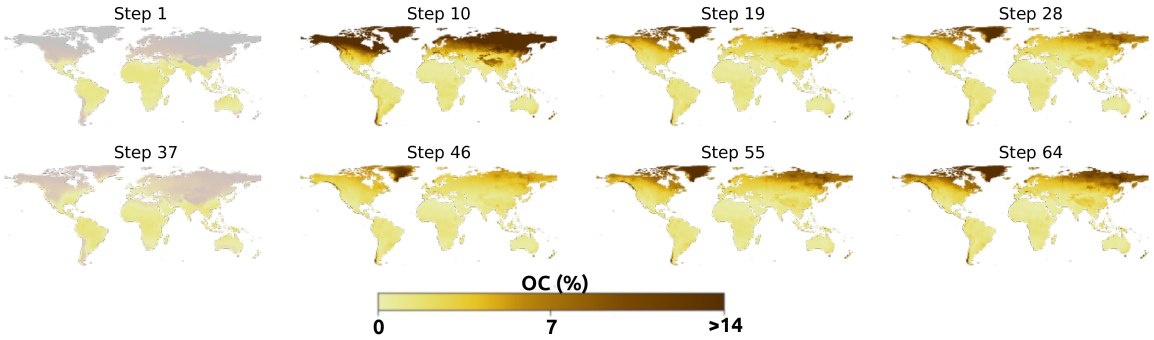


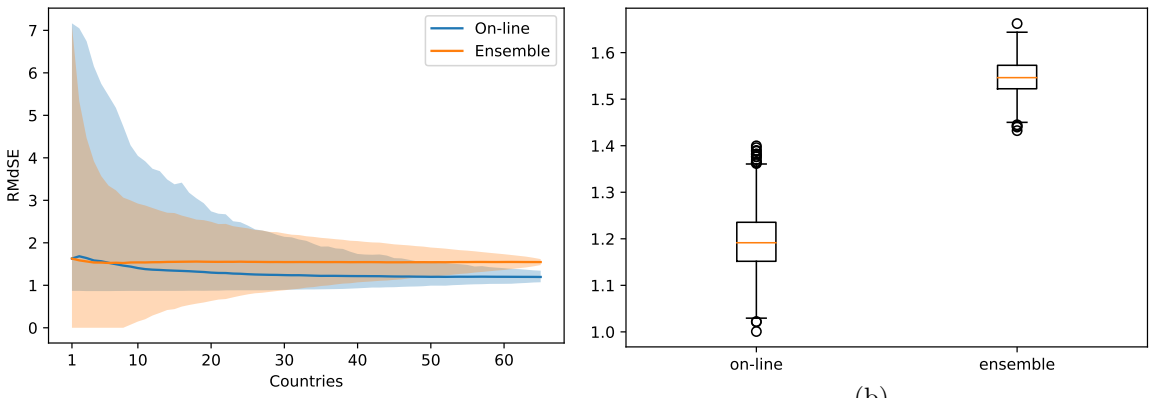
Fig. 3.5: Sequence of maps at different training steps. Maps correspond, from top-left to bottom right, to steps 1, 10, 19, 28, 37, 46, 55, and 64. Note that the maps were generated during a single iteration (not 1,000 repetitions) of the online learner.

3.4.2 Online versus ensemble

Similar to the online model, the ensemble model, which averages outputs from the model trained at each country, shows a decrease in the error variance as models from different countries are added to the model (Fig. 3.6). However, the error from the model ensemble shows a larger error than the online learning model. The error variance reduction due to model averaging is a common phenomenon exploited in ML, where bootstrap aggregation (bagging) (Breiman, 1996) is one of the simplest and preferred methods. The error of the model ensemble slightly increases, which is a common result of bagging as pointed out by Zhang (2004).

For simplicity, at the moment of selecting an averaging method, we used a simple equal weight average since more sophisticated methods are usually dependent on statistics not easily available under distributed training conditions. For instance, the methods proposed by Granger and Ramanathan (1984), the weights correspond to the ordinary least squares estimates of a multiple linear regression that depends on a vector of the observed values. Once a practical validation strategy is generated (see Section 3.4.5) it would be possible to evaluate other ensemble methods which might perform better than the equal weight average.

Regarding computational time, both online and ensemble models performed similarly. However, the time for model prediction linearly increases with the number of models used in the ensemble. This could be an issue and could potentially be crucial in some domains, for instance Digital Soil Mapping where the combination of



(a) (b)

Fig. 3.6: (a) Mean RMdSE (% SOC) for each consecutive training. Shaded areas correspond to the the difference between the 97.5 and 2.5 percentiles; (b) RMdSE (% SOC) after the last country is added to the model. The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR). All the calculations are based on 1,000 iterations.

large extents, high resolution, and bootstrap routines, leads to running a single model realisation on billions of pixels. This computational burden can take several hours or days to generate a model output (Padarian *et al.*, 2014; Padarian *et al.*, 2017; Hengl *et al.*, 2017).

3.4.3 Response to redundant data

As already mentioned, in a real world application, a platform implementing this algorithm will most likely have to deal with redundant information (same piece of data is held in two separate places). As observed in Fig. 3.7, the incremental inclusion of already-seen data could have a detrimental effect on the learning process. It is possible to observe a shift of the final prediction error, and, more importantly, a significant increase in the variance of the predictions.

As we aim to continuously add more information to the model, it is also important to avoid feeding redundant information (Japkowicz and Stephen, 2002), which is similar to the data duplication problem (having multiple copies of some samples, generally by mistake) as described by Kołcz *et al.*, (2003). Imbalanced data is a common problem in

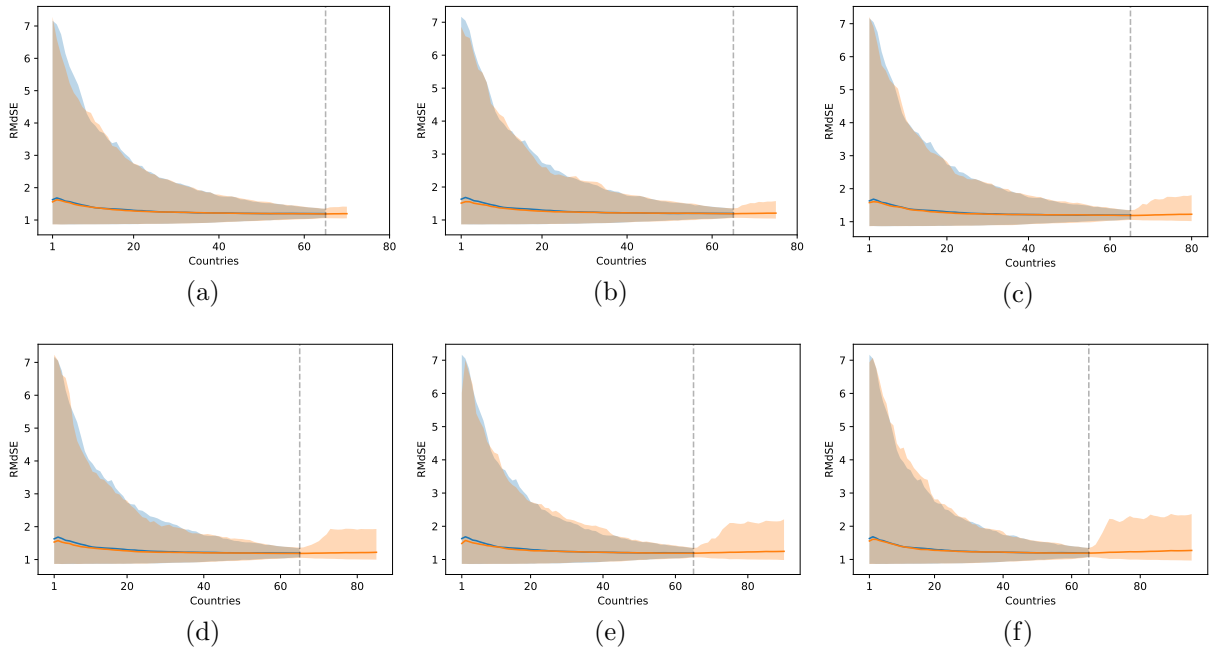


Fig. 3.7: Error (% SOC) and variance evolution after including redundant data. (a) 5 already-seen countries; (b) 5 already-seen countries repeated 2 times; (c) 5 already-seen countries repeated 3 times; (d) 5 already-seen countries repeated 4 times; (e) 5 already-seen countries repeated 5 times; (f) 5 already-seen countries repeated 6 times. Vertical dashed line delimits the beginning of the redundant data inclusion.

ML (Chawla *et al.*, 2004), which is known to reduce the performance of the algorithms. In our case study, the increase in the variance implies a reversion of the process depicted in Fig. 3.4, shifting from a global model to a local one. This problem needs to be tackled when we aim to create a collaborative global model. One possible solution is to apply weights to new data so data that are already in the domain of prediction will be weighted less.

3.4.4 Hyperparameter selection

Hyperparameters are the group of parameters of a ML algorithm which cannot be learned in the training process (e.g., learning rate, batch size, etc.). Neural networks are very sensitive to these hyperparameters, which in some cases can increase the computation time, and, more importantly, reduce the accuracy or even cause the

algorithm never to converge. Hyperparameter selection can be performed in different ways, from an exhaustive grid search, where all the combinations of parameters are tested, to random selection from a range of values (Bergstra and Bengio, 2012) or from a distribution (Snoek *et al.*, 2012).

In large scale online learning, the optimisation is usually performed from a small sub-sample of the data which yields good results (Bottou, 2012). This condition is only true when the sub-sample is representative of the data, which poses a challenge in situations like our case study, where at each step we incorporate data from different countries or places, and the initiator party (first party or country) does not have access to data from other countries. Thus an initial training from public domain data could help in setting the best hyperparameters. It is advisable to follow some general advice about setting hyperparameters in the context of online learners and SGD. For more detail on this subject, we refer the reader to Bottou (2012).

3.4.5 Validation strategy

During this work, we were able to validate the model at each step during the training because we had access to all the data and we simulated the situation were the data was distributed among different countries. In a real application, the validation strategy will most likely vary. Similar to hyperparameter selection, it would be possible to use a public domain dataset with similar characteristics to the expected final model coverage in order to validate each step of the training.

3.4.6 Platform

In order to implement our approach, we envision the use of a platform to facilitate collaborative modelling, and model sharing. Similar work has been done by the project Berkeley Open Infrastructure for Network Computing (Anderson, 2004), where data is transferred to different computers to be processed. In our case, instead of transferring data, the model will "travel" to agencies that have agreed to collaborate. The agency will train the model using its data and only the model parameters will be sent to the main platform.

The proposed platform addressed some of the technical issues described previously which can be summarised below.

Data redundancy: Keeping track of trained data by only storing metadata which includes data domain information. For instance, define the geographic area covered by the observations or statistical distribution of the covariates. Data redundancy can be addressed by a variation of data declustering in the geographical space (Journel, 1983) or, alternatively, relative to the covariate space (Carré *et al.*, 2007). Data thus can be weighted accordingly in the model training, avoiding over-training on certain parts of the geographical or covariate space.

Hyperparameter selection and validation: As an alternative to using global public domain data, if multiple datasets are simultaneously available in the platform (each held by their corresponding agency), and, assuming that the combination of them is more representative than the initiation dataset, it would be technically possible to perform a better hyperparameter optimisation and validation.

3.5 Conclusions

This work addresses the problem of collaborative modelling when data privacy is a concern and restricts data sharing. We successfully emulated a situation where different countries provide soil data (without sharing it) and, by using an online learning algorithm trained one dataset at a time, we showed that is possible to yield a similar prediction accuracy compared with the model that is trained with all the data.

We believe that, even if in some cases the accuracy can be affected, it is possible to compromise on the accuracy loss with the access to more information which provides a broader range of conditions from where a machine learning algorithm can generate its rules.

The proposed method for collaboration learning tackles the data privacy concerns which frequently hampers global collaboration. The proposed online learning method is unique and distinct from privacy preserving data mining algorithms which require data encryption. The method utilised the optimisation algorithm meaning that it can be used in a variety of ML models, for classification or regression tasks.

3.6 References

Anderson, D. P. (2004). Boinc: A system for public-resource computing and storage. In: *Grid Computing, 2004. Proceedings. Fifth IEEE/ACM International Workshop on*. IEEE: pp. 4–10.

Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., *et al.*, (2017). Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ* 14: 1–19.

Batjes, N. (2009). Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. *Soil Use and Management* 25 (2): 124–127.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13 (Feb): 281–305.

Birge, R. T. (1939). The propagation of errors. *American Journal of Physics* 7 (6): 351–357.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63 (6): 1059–1078.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks* 17 (9): 142.

Bottou, L. (2012). Stochastic gradient descent tricks. In: *Neural networks: Tricks of the trade*. Springer: pp. 421–436.

Bousquet, O. and Bottou, L. (2008). The tradeoffs of large scale learning. In: *Advances in neural information processing systems*: pp. 161–168.

Breiman, L. (1996). Bagging predictors. *Machine learning* 24 (2): 123–140.

Bulatewicz, T., Allen, A., Peterson, J. M., Staggenborg, S., Welch, S. M., and Steward, D. R. (2013). The simple script wrapper for OpenMI: enabling interdisciplinary modeling studies. *Environmental Modelling & Software* 39: 283–294.

Carré, F, McBratney, A. B., and Minasny, B (2007). Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma* 141 (1-2): 1–14.

Cauchy, A. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris* 25 (1847): 536–538.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter* 6 (1): 1–6.

CIAT (2013). *Creación del Sistema de Información de Suelos de Latinoamérica (SISLAC). Fase I. Final report*. Tech. rep. Updated 2014.

Danielson, J. J. and Gesch, D. B. (2011). *Global multi-resolution terrain elevation data 2010 (GMTED2010)*. Tech. rep. US Geological Survey.

DeSantis, T., Dubosarskiy, I, Murray, S., and Andersen, G. L. (2003). Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. *Bioinformatics* 19 (12): 1461–1468.

Dolk, D. R. and Kottemann, J. E. (1993). Model integration and a theory of models. *Decision Support Systems* 9 (1): 51–63.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Vol. 57. New York: CRC press: p. 436.

Fienberg, S. E., Martin, M. E., Straf, M. L., Council, N. R., *et al.*, (1985). *Sharing research data*. National Academies.

Granger, C. W. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of forecasting* 3 (2): 197–204.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., *et al.*, (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PloS one* 12 (2): e0169748.

Hiederer, R (2010). Data update and model revision for soil profile analytical database of Europe of measured parameters (SPADE/M2). *JRC Scientific and Technical Reports, EUR* 24333.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* 25 (15): 1965–1978.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis* 6 (5): 429–449.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology* 15 (3): 445–468.

Kantarcioglu, M. and Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE transactions on knowledge and data engineering* 16 (9): 1026–1037.

Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G., and de Vries, F. (2012). Efficiency comparison of conventional and digital soil mapping for updating soil maps. *Soil Science Society of America Journal* 76 (6): 2097–2115.

Kołcz, A., Chowdhury, A., and Alspector, J. (2003). Data duplication: An imbalance problem?

Lee, M., hong Jeon, J., Kim, J., and Song, J. (2010). Scalable and parallel implementation of a financial application on a GPU: With focus on out-of-core case. In: *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*. IEEE: pp. 1323–1327.

Maxwell, T. (1999). A paris-model approach to modular simulation. *Environmental modelling & software* 14 (6): 511–517.

McBratney, A., Mendonça Santos, M. L., and Minasny, B (2003). On digital soil mapping. *Geoderma* 117 (1): 3–52.

Minasny, B., Malone, B. P., Mcbratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B. S., *et al.*, (2017). Soil carbon 4 per mille. *Geoderma* 292: 59–86.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*: pp. 807–814.

Padarian, J, Minasny, B, and McBratney, A. B. (2014). The Evolving Methodology for Global Soil Mapping. In: *GlobalSoilMap: Basis of the global spatial soil information system*. Ed. by D Arrouays, N McKenzie, J Hempel, A. Richer de Forges, and A. B. McBratney. London, UK: Taylor & Francis Group: pp. 215–220.

Padarian, J, Minasny, B, and McBratney, A. (2017). Chile and the Chilean soil grid: a contribution to GlobalSoilMap. *Geoderma Regional*: 17–28.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Piwowar, H. A., Day, R. S., and Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS one* 2 (3): e308.

Poelen, J. H., Simons, J. D., and Mungall, C. J. (2014). Global biotic interactions: an open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics* 24: 148–159.

Sarle, W. S. (1994). Neural networks and statistical models.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In: *Advances in neural information processing systems*: pp. 2951–2959.

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., Hong, S. Y., Rawlins, B. G., and Field, D. J. (2015). Global soil organic carbon assessment. *Global Food Security* 6: 9–16.

Stockwell, D., Davey, S., Davis, J., Noble, I., *et al.*, (1990). Using induction of decision trees to predict greater glider density. *AI Applications in Natural Resource Management* 4 (4): 33–43.

Toledo, S. (1999). A survey of out-of-core algorithms in numerical linear algebra. *External Memory Algorithms and Visualization* 50: 161–179.

Wieland, S. C., Cassa, C. A., Mandl, K. D., and Berger, B. (2008). Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Sciences* 105 (46): 17608–17613.

Yost, R., Uehara, G., Wade, M., Sudjadi, M, Widjaja-Adhi, I., and Zhi-Cheng, L. (1988). Expert systems in agriculture: Determining lime recommendations for soils of the humid tropics.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: *Proceedings of the twenty-first international conference on Machine learning*. ACM: p. 116.

Chapter 4

Using deep learning to predict soil properties from regional spectral data

Summary

Diffuse reflectance infrared spectroscopy allows the rapid acquisition of soil information in the field or the laboratory. The vis-NIR spectroscopy research enthusiasm around the world has created regional to global soil spectral libraries. While machine learning methods have been utilised in processing spectral data, such large regional datasets are better dealt with by big data analytics. Deep learning is an exciting discipline that has already transformed the way data are analysed in many fields and could also change the way we model soil spectral data. This study evaluated convolutional neural networks (CNNs), a type of deep learning algorithm, as a new way to predict soil properties from raw soil spectra. We demonstrated the effectiveness of CNNs on the LUCAS soil database, which consists of around 20,000 topsoil observations with physicochemical and biological properties from Europe. To fully utilise the capacity of the CNN model, we represented the soil spectral data as a 2-dimensional spectrogram, showing the reflectance as a function of wavelength and frequency. We showed the capacity of a CNN to be trained in a multi-task setting to simultaneously predict six soil properties in one model (organic carbon, clay, sand and total nitrogen content, cation exchange

capacity, and pH). Compared with conventional methods such as partial least squares (PLS) regression and Cubist regression tree, the CNN performed significantly better, especially the multi-tasking model. In the case of soil organic carbon prediction, the multi-task CNN decreased the error on a held-out dataset by 87% compared to PLS and 62% compared with Cubist. This approach proved to be effective when trained on a relatively large dataset.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. there are no other authors of the publication according to these criteria;

4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and

5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for this publication is:

Padarian, J., Minasny, B. and McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. Geoderma Regional, 16, p.e00198.

Contributors	Statement of contribution
José Padarian *Signature:* José Padarian *Date:* April 2, 2020	Conceptualisation Data analysis Writing
Budiman Minasny	Writing
Alex McBratney	Writing

4.1 Introduction

Diffuse reflectance infrared spectroscopy, in both the visible-near (VNIR, 350–2500 nm) and mid-infrared wavelength ranges (MIR, 2500–25000 nm), allows rapid acquisition of soil information in the field or in the laboratory. In particular, VNIR spectroscopy has been studied extensively for the rapid prediction of soil properties. Soil spectra in the VNIR region have been found to provide good estimates for a large range of soil physical, chemical and biological properties. Excellent results were reported for measurement of total carbon, total nitrogen, clay and sand content, cation exchange capacity (CEC), pH, and microbial activity (Bellon-Maurel and McBratney, 2011; Soriano-Disla *et al.*, 2014). VNIR spectrometry has gained popularity in soil science as it is available in portable instruments, is easy and ready to use in the field and requires minimal or even no sample preparation (Viscarra Rossel *et al.*, 2009). Comprehensive reviews on the use of VNIR for predicting soil properties can be found in Stenberg *et al.*, (2010) and Soriano-Disla *et al.*, (2014).

The VNIR enthusiasm around the world has created country (e.g. (Romero *et al.*, 2018)), regional (Shepherd and Walsh, 2002), continental (Stevens *et al.*, 2013), and global (Viscarra Rossel *et al.*, 2016) spectral libraries, with the desire that such big data can fully describe soil composition. However, soil properties are covertly encoded and we need to extract useful information from the spectral data to be able to predict them. Since soil is a complex mixture of materials, it is sometimes difficult to assign specific features of the spectra to specific chemical components. In addition, ultraspectral data obtained from spectrometers contain thousands of reflectance values as a function of wavelength. As there are usually more predictor variables than observations and predicted soil attributes, methods that reduce the dimension of the spectra such as partial least squares (PLS) regression (Martens and Naes, 1989) are commonly used. PLS regression extracts successive linear combinations of the spectra, which optimally address the combined goals of explaining response variation and explaining predictor variation. This is the most common or standard practice in chemometrics and soil spectroscopy. Machine learning techniques that are capable of handling large amounts of input variables (e.g. support vector machine, artificial neural networks, random forests) have also been tested for calibrating soil spectra (Morellos *et al.*, 2016; Stevens *et al.*, 2013; Viscarra Rossel and Behrens, 2010). Other models that

include variable (wavelength) selection have also been found useful (Minasny and McBratney, 2008; Sarathjith *et al.*, 2016). In addition to the dimensionality problem, spectra pre-processing affects prediction accuracy (Gras *et al.*, 2014; Vašát *et al.*, 2017). Smoothing spectra with methods such as the Savitzky-Golay polynomial smoothing, and standardising spectra via Standard Normal Variate (Stevens *et al.*, 2013) are common practices. Spectra sampling or compression to reduce the excessive number of predictors are also commonly performed (Viscarra Rossel *et al.*, 2016).

With the development of large spectral libraries, we need to seize the opportunity to utilise big data analytics to help use and process the spectral data which goes beyond using commercial software or packaged machine learning algorithms. Currently, deep learning is a rapidly developing frontier in machine learning that has been widely used in image and speech recognition (LeCun *et al.*, 2015) thanks to their ability to learn multiple (hierarchical) levels of data representation (Bengio, 2012). This development is enhanced with the availability of big data, computing power, new algorithms and numerical computational tools such as TensorFlow (Abadi *et al.*, 2015). Deep learning algorithms have been used to classify and extract features of hyperspectral data (Chen *et al.*, 2014) but, as far as we are aware, they have not been used in soil VNIR spectroscopy prediction.

The objective of this work is to explore the use of deep learning, specifically convolutional neural networks (CNNs), to predict soil properties from unprocessed soil spectral data, avoiding dimension reduction and pre-processing procedures. First, we introduce CNNs, explaining how they work internally and how they extract features from the data. Second, we introduce the use of spectrograms as a better way to represent spectral data for model generation. Third, we explore a multi-task approach where we predict multiple soil properties simultaneously with a single model. Finally, we compare the performance of our proposed method with prediction methods commonly used in soil spectroscopy (PLS and Cubist), and we evaluate the effectiveness of our approach when facing datasets of different size.

4.2 Convolutional neural networks

Deep learning is a model that is made of multiple processing layers to learn data representation (LeCun *et al.*, 2015). Deep learning is different from traditional neural

networks, which have been used in soil spectra processing, as it involves more layers and deeper architectures. Deep neural networks allow the use of raw or unprocessed data (e.g. images or spectra) and automatically discover the representations needed for prediction. The data are transformed at each layer, amplifying aspects of the input data that are important and suppressing irrelevant information for enhanced prediction.

One such deep learning model is the convolutional neural network (CNN), which is a neural network that includes one or more convolutional layers in its architecture. CNN is designed to take data in the form of multiple arrays such as images. A convolutional layer performs convolutions over an array (Fig. 4.1) using multiple filters. These layers are connected by weights, which are learned during training. After training, each filter can identify different features, similar to an image edge detection filter (e.g.: Sobel-Feldman; Eq. 4.1). A single convolutional layer is capable of identifying simple features and, as more layers are added, the network is capable of extracting features of increasing complexity and abstraction (LeCun *et al.*, 1990).

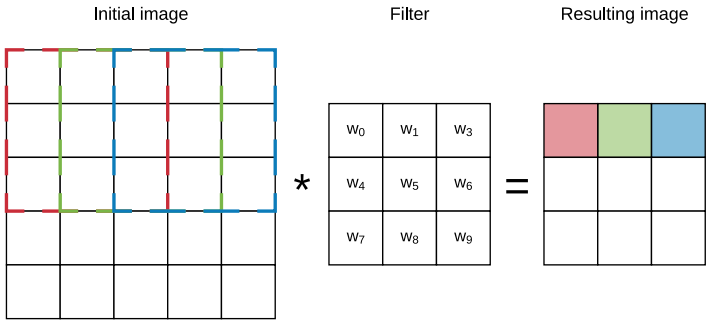


Fig. 4.1: Example of the first 3 steps of a convolution of a 3x3 filter over a 5x5 array (image). The resulting pixel values correspond to the sum of the element-wise multiplication of the initial pixels (dashed lines) and the filter.

$$S_{horizontal} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}; \qquad S_{vertical} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{4.1}$$

CNNs have been widely used in computer vision, but also in speech recognition and time-series analysis (LeCun and Bengio, 1995). They have the capacity of exploiting local correlations (spatial, temporal or both, depending on the input data) which makes them more suitable for this type of analysis than traditional fully-connected neural

networks. For more insights about CNN we refer the reader to the seminal works of LeCun *et al.*, (1990) and Krizhevsky *et al.*, (2012).

4.3 Spectrograms

Spectrograms appeared in signal processing to visualise a sound as early as 1947, with one of the first examples shown by Potter *et al.*, (1947). Spectrograms are a representation of a signal in a 2D space (e.g.: time-frequency for audio), where the magnitude of the signal is represented by the value of the pixels. Spectrograms are usually generated by decomposing the signal into overlapping segments to which a short-time fast Fourier transformation is applied (Griffin and Lim, 1984).

To generate the spectrograms, we used a Hann window (Blackman and Tukey, 1958), a segment length of 100, with 50% of overlap, and a sampling frequency of one (i.e. not skipping data). After generating the spectrograms, they were transformed to a logarithmic scale. By processing the spectral data in this way, we generated a different (2D) representation of the spectrum which is more appropriate for our CNN (based on 2D convolutions) to process, from a vector of length 4,200 to a matrix of 51x83 (frequency x wavelength). Examples of the original data and the resulting spectrograms are shown in Fig. 4.2.

In this new representation, it is still possible to distinguish the drops in reflectance as larger (brighter) values in a short window and wide range of frequencies, and also a subtle overall increase of values (a "glow") at lower levels of reflectance (Fig. 4.2, right panels). In effect, spectrograms automatically performed and represented multiple scaling of the spectra as in wavelet transform.

4.4 CNN Model

Designing a CNN is a highly iterative process which includes making decisions on hyperparameter such as the number and type of layers used, and learning rate. Here we introduce a CNN model for processing spectral data and present the networks with the best performing combination of hyperparameters.

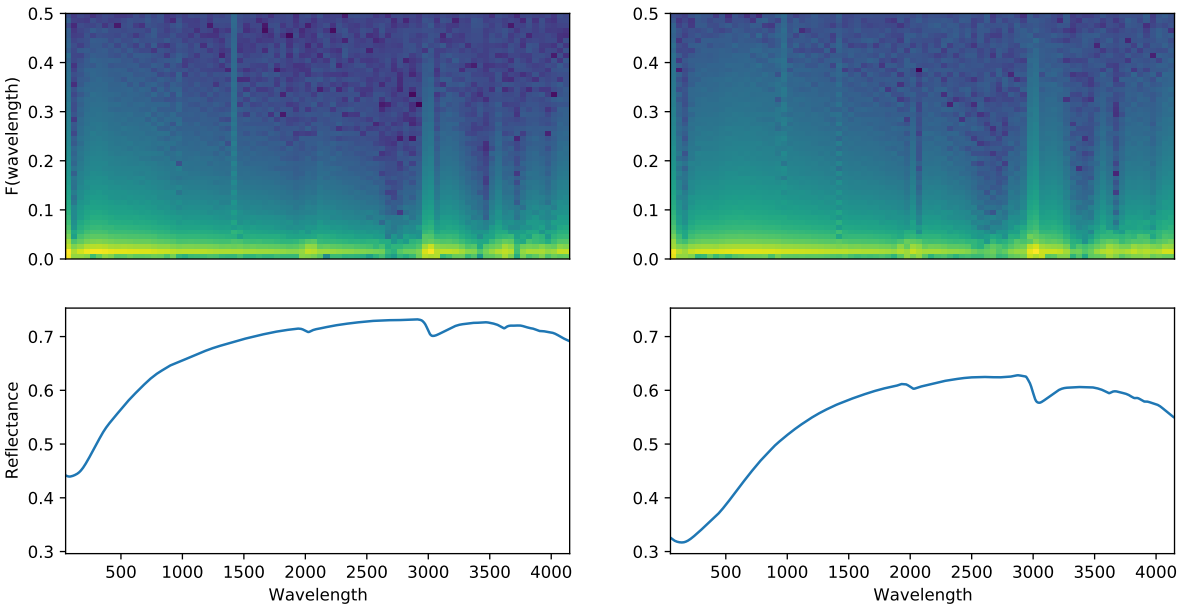


Fig. 4.2: Example of spectral data encoded as a spectrogram. Top panels: spectrogram with amplitude (colour) in log scale. Bottom panels: original spectral data. Left panels: mineral soil (0.5% organic carbon). Right panels: organic soil (20% organic carbon)

4.4.1 Network architecture

The CNN follow a series of combinations of convolution and pooling layers (Table 4.1). Units in a convolutional layer are organised in feature maps (here we used 3x3). A feature map, represented as a square in Fig. 4.1, corresponds to the output for one of the learned features which are detected at each of the image positions. Each unit of the feature map is connected to local patches in the feature maps of the previous layer through a set of weights. These locally weighted sums are then passed through a non-linear function, such as the ReLU (Rectified Learned Unit) function: $f(x) = max(0, x)$, where x is the input to a neuron.

Following the convolution layer, Max-Pooling layers combine inputs from the convolutional layers using a 2x2 window, thus reducing the resolution of the feature maps (Scherer *et al.*, 2010). After several stages of convolution and pooling, the results are flattened (to a 1D array) and followed by fully-connected layers. These convolutional and pooling layers are based on the principles of a network mixing simple and complex cells in neuroscience. CNN takes into account the compositional hierarchies of data, e.g., local signals of spectra form peaks and valleys, and the whole

forms a spectrum. Thus the idea is to extract a representation from local information at various scales to predict soil properties.

As the information moves through the network, the representation changes (Fig. 4.3), going from an 83x51 image to the single value prediction. Table 4.1 summarises the architecture of the CNN used in this study. It contains seven trainable layers — five convolutional and two fully-connected layers.

Table 4.1: Sequence of layers used to build the neural network.

Type	Kernel size	Filters	Activation
Convolutional	3x3	64	ReLU
Max-Pooling	2x2	-	-
Convolutional	3x3	128	ReLU
Convolutional	3x3	256	ReLU
Max-Pooling	2x2	-	-
Convolutional	3x3	512	ReLU
Convolutional	3x3	512	ReLU
Fully-connected	-	100	ReLU
Fully-connected	-	1	Linear

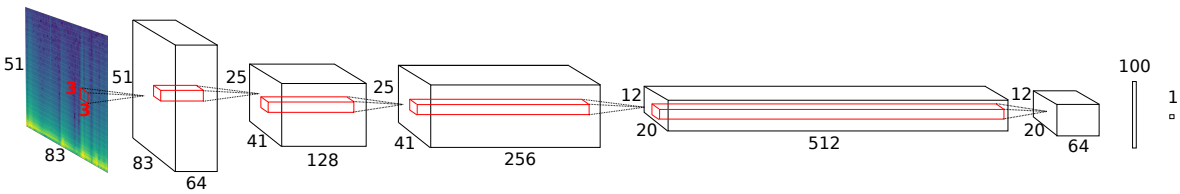


Fig. 4.3: Sequence of layers showing the information flow from an input spectrogram (left end) to a single value prediction (right end).

4.4.2 Multi-task network

CNNs have the capacity to predict multiple soil properties in a single network and training process. To utilise this capability, we used the same six soil properties and we varied the network architecture (Table 4.2). The new architecture has a series of four shared convolutional layers, followed by, for each property, a series of one convolutional and one fully-connected layer (Fig. 4.4).

Table 4.2: Sequence of layers used to build the multi-task neural network

Layer type	Kernel size	Filters	Activation
†Convolutional	3x3	64	ReLU
†Max-Pooling	2x2	-	-
†Convolutional	3x3	128	ReLU
†Convolutional	3x3	256	ReLU
†Max-Pooling	2x2	-	-
†Convolutional	3x3	512	ReLU
‡Bottle-neck	1x1	64	ReLU
‡Fully-connected	-	1	Linear

†Common layers; ‡for each property

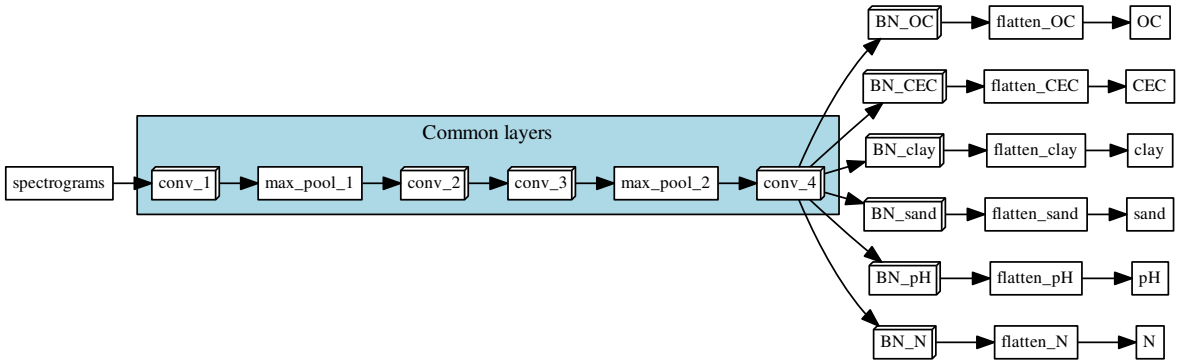


Fig. 4.4: Architecture of the multi-task network. "Common layers" represent the layers shared by all the predicted properties. Each branch, one per predicted soil property, correspond to a series of one convolutional layer (BN: bottle-neck layer, which reduces the dimensionality of the data) and a fully-connected layer of size=1, which corresponds to the final prediction

The head of the network ("Common layers") is a series of convolutional and Max-Pooling layers. This section of the network is shared by all the target soil properties and should be able to learn how the spectrogram is structured. After the "Common layers" extract a general representation of the data represented by the spectrogram, the information is directed to 6 different branches, one for each target soil property. Each branch consists of a convolutional layer (BN) which is flattened (to 1D) before generating the output . The branches should be able to learn the signals found in the spectrogram that are specific for each soil property.

The effect of using a multi-task model has been widely studied. In a review by Ruder

(2017), it is mentioned that using a multi-task model reduces the risk of overfitting (by using additional information, acting as a regularisation) and, more notably, the accuracy increases continuously with the number of tasks (Ramsundar *et al.*, 2015).

4.4.3 Training the network

To find the optimal weights of the network, the network needs to be trained using the data. Usually, when using CNNs the data is processed in batches, which allows the use of big datasets that cannot fit in memory at once. One cycle through the entire training dataset is a training epoch. During the training, the weights are adjusted based on a gradient-based optimisation method, meaning that the partial derivatives of the parameters with respect to the error were evaluated and the parameters were adjusted towards a minimum error value. The rate of change of parameters along the error gradient is controlled by the learning rate. If the learning rate is too high, the weights will change far too much with each iteration, which will make the parameters "bounce" around the optimal solution, or simply diverge. If the learning rate is too low, the parameters might never converge.

In this work, we trained the network during 20 epochs, using a batch size of 10 and the *Adam* optimiser (Kingma and Ba, 2014). *Adam* maintains a learning rate for each parameter and modifies them during the training based on exponential decays, β_1 and β_2, of the first- and second-moment of the estimates respectively.

4.4.4 The data

We used 2 datasets to test the performance of CNN. The first is a large regional dataset (LUCAS) from Europe with 19,036 topsoil (composite of five sub-samples of the top 30 cm) observations with physicochemical and biological properties from Europe (Stevens *et al.*, 2013). Samples are distributed over all land use/cover types, with some over-representation of croplands and under-representation of Eastern Europe. The soil samples were ground and passed through a 2 mm sieve. The samples were scanned with a diffuse reflectance spectrometer (XDS$^{\text{TM}}$ Rapid Content Analyzer) with reflectance data from 400 to 2500 nm at a spectral resolution of 0.5 nm, resulting in 4200 wavelengths.

From the LUCAS dataset, we selected six soil properties which were known to be well predicted by the VNIR spectra: *a)* organic carbon content (OC, g kg^{-1}), *b)* cation exchange capacity ($cmol^{+}$ kg^{-1}), *c)* clay particle size fraction (%), *d)* sand particle size fraction (%), *e)* pH measured in water, and *f)* total nitrogen content (N, g kg^{-1}). A summary of statistics of these soil properties is presented in Table 4.3.

Table 4.3: Summary statistics of soil properties (n=19,036) for the LUCAS Soil database (Stevens *et al.*, 2013).

	OC ($g\ kg^{-1}$)	CEC ($cmol^{+}\ kg^{-1}$)	Clay (%)	Sand (%)	pH	N ($g\ kg^{-1}$)
Minimum	0.00	0.00	0.00	1.00	3.21	0.00
Maximum	586.80	234.00	79.00	99.00	10.08	38.60
Mean	50.00	15.76	18.88	42.88	6.20	2.92
Median	20.80	12.40	17.00	42.00	6.21	1.70
St. Dev.	91.31	14.48	13.00	26.11	1.35	3.76
Skewness	3.67	4.24	0.91	0.19	-0.08	3.76

To test the effect of dataset size, we used a second, smaller, dataset from the study of Geeves *et al.*, (1995). The dataset represents 72 soil profiles (390 samples) in the wheat-belt of southern NSW and northern Victoria, Australia. The samples covered a range of soil types and were taken from different soil horizons up to 1 m depth. The soil samples were ground and passed through a 2 mm sieve. Samples were scanned using an AgriSpec^{TM} instrument for reflectance spectra (350–2500 nm, 1nm resolution). We selected five soil properties, namely total carbon (g kg^{-1}), CEC, clay and sand content, and pH (in $CaCl_2$). A summary of statistics of soil properties for these soil samples is presented in Table 4.4.

For the LUCAS database, we modelled organic and mineral samples together as we tried to model only using spectral data without having any prior information about the samples. The soil properties and spectra of both datasets were used here for testing the CNN model.

4.4.5 Training & Validation

In this study, we divided the LUCAS dataset into a training set, a validation set, and a test set. The training set is used to fit or train the models; the validation set is used

Table 4.4: Summary statistics of soil properties for the dataset by Geeves *et al.*, (1995) (n=390).

	Total Carbon ($g\ kg^{-1}$)	CEC ($cmol^{+}\ kg^{-1}$)	Clay (%)	Sand (%)	pH
Minimum	0.06	0.40	5.00	14.00	3.76
Maximum	12.74	36.43	74.00	91.00	8.23
Mean	1.17	9.17	26.60	56.92	5.49
Median	0.85	7.34	20.00	60.00	5.32
St. Dev.	1.37	5.79	16.62	16.50	0.98
Skewness	4.22	1.40	1.02	-0.53	0.67

to estimate prediction error for parameter selection; and the test set is used to assess the error of the model. From the full dataset, 25% (n = 4,759) of the samples were randomly selected and used as a test set. The rest of the data were used in training and validation. We performed a bootstrapping routine (Efron and Tibshirani, 1993) with 100 repetitions and measured the accuracy of a prediction by generating different models from different realisations of the data. A bootstrapping routine assumes that the training data set is a representation of the population, and multiple realisations of the population can be simulated from a single dataset. This is done by repeated random "sampling with replacement" of the original dataset of size 14,277 to obtain 100 bootstraps, each of size 14,277. Theoretically, about 2/3 of the data will be used in training in a bootstrap iteration, and the remaining 1/3 of data is used as validation. And thus we have a 50:25:25 percent split of data into training, validation, and test sets.

For all datasets, and at each bootstrap iteration, we estimated the root mean squared error (RMSE), coefficient of determination (R^2), mean error (ME) of prediction, and concordance correlation coefficient (ρ_c; Lawrence and Lin, 1989).

To compare our results with conventional techniques, we generated predictive models using a Cubist regression tree (Quinlan, 1992) and Partial Least Squares regression (PLS; Martens and Naes, 1989) model, which are commonly used in soil spectroscopy studies (Brown *et al.*, 2006; Minasny and McBratney, 2008; Stevens *et al.*, 2013; Niazi *et al.*, 2015; Viscarra Rossel *et al.*, 2016). Before training the Cubist and PLS models, the spectral data were pre-processed using a series of methods commonly

used in the literature: *a)* converting reflectance to apparent absorbance ($a = -log_{10}(r)$); *b)* Savitzky–Golay smoothing (Savitzky and Golay, 1964), using a window size of 11, and a second order polynomial; *c)* edges trimming (< 500 nm and > 2450 nm) to discard noisy data; *d)* sampling every tenth measurement; and *e)* applying a standard normal variate transformation (Barnes *et al.*, 1989).

For the small dataset (Geeves *et al.*, 1995), the data were randomly split into training/validation sets during the bootstrapping routine. Considering the small number of samples, we did not perform a training/validation/testing split. We used a similar network like the one described in Table 4.1 but with some minor modifications to avoid overfitting. We replaced the first fully-connected layer with a dropout rate (Nitish *et al.*, 2014) with a probability of 0.5. As a reference, we also compared the results with the Cubist model.

4.4.6 Implementation

The CNN was implemented in Python (v3.6.2; Python Software Foundation, 2017) using Keras (v2.1.2; Chollet, 2015) and Tensorflow (v1.4.1; Abadi *et al.*, 2015) backend. The Cubist and PLS models were implemented in R (v3.3.1; R Core Team, 2016), using the packages Cubist (v0.2.1; Kuhn and Quinlan, 2017) and pls (v2.6-0; Mevik *et al.*, 2016) respectively. Computing was done using the University of Sydney's Artemis high performance computing facility.

4.5 Results and discussion

4.5.1 Training

To evaluate the effectiveness of the CNN model in predicting soil properties, we listed the performance of the model (Table 4.5) which shows good results, with mean R^2 ranging from 0.63 for sand to 0.94 for OC. Not surprisingly, the performance in the training datasets is better than on the test and validation datasets. The models do not overfit, as shown by the similar performance between the training and validation datasets. The similarity in the error of training, validation, and test sets is also a sign of good hyperparameter selection. The estimates are slightly biased, however, the bias

is generally less than 10% of the minimum values of the properties (Table 4.3).

Table 4.5: Training statistics using multi-task CNN for OC (g kg^{-1}), CEC ($cmol^{+}$ kg^{-1}), clay content (%), sand content (%), pH and N (g kg^{-1}). Mean, standard deviation (sd), minimum (min) and maximum (max) of 100 bootstrap realisations.

		Train (n = 50%)			*Validation* (n = 25%)			*Test* (n = 25%)		
		RMSE	R^2	ME	RMSE	R^2	ME	RMSE	R^2	ME
OC	mean	24.75	0.94	-0.79	29.78	0.90	-0.45	28.83	0.90	-1.29
	sd	2.66	0.01	1.12	1.82	0.01	1.10	1.63	0.01	1.40
	min	20.26	0.93	-1.58	27.73	0.89	-1.23	26.93	0.89	-2.28
	max	28.25	0.96	0.00	33.38	0.91	0.32	31.86	0.91	-0.31
CEC	mean	7.75	0.74	2.34	8.52	0.66	2.21	8.68	0.65	2.10
	sd	0.40	0.03	0.26	0.21	0.02	0.15	0.47	0.02	0.33
	min	7.29	0.68	2.16	8.18	0.63	2.11	7.97	0.62	1.86
	max	8.32	0.77	2.53	8.84	0.70	2.32	9.65	0.68	2.33
Clay	mean	6.46	0.77	-0.53	7.37	0.69	-0.61	7.47	0.68	-0.65
	sd	0.53	0.03	2.75	0.34	0.02	2.94	0.26	0.02	2.86
	min	5.65	0.72	-2.47	6.78	0.67	-2.69	7.06	0.65	-2.67
	max	7.07	0.81	1.41	7.85	0.73	1.47	7.78	0.70	1.37
Sand	mean	15.86	0.63	-1.53	17.85	0.54	-1.30	18.03	0.54	-1.16
	sd	1.75	0.08	0.61	0.65	0.03	0.51	0.59	0.02	0.60
	min	12.09	0.53	-1.96	17.01	0.50	-1.67	17.21	0.50	-1.58
	max	17.97	0.80	-1.10	18.75	0.57	-0.94	19.01	0.57	-0.73
pH	mean	0.44	0.90	0.04	0.49	0.87	0.04	0.50	0.87	0.04
	sd	0.02	0.01	0.02	0.01	0.00	0.02	0.01	0.00	0.02
	min	0.42	0.89	0.02	0.48	0.87	0.03	0.49	0.86	0.02
	max	0.46	0.91	0.05	0.52	0.88	0.05	0.52	0.87	0.06
N	mean	1.30	0.89	-0.19	1.51	0.85	-0.21	1.52	0.83	-0.22
	sd	0.05	0.01	0.18	0.06	0.01	0.14	0.04	0.01	0.17
	min	1.22	0.88	-0.32	1.45	0.84	-0.31	1.47	0.81	-0.34
	max	1.38	0.90	-0.07	1.63	0.86	-0.11	1.59	0.85	-0.10

The observed errors are comparable with the results reported by Stevens *et al.*, (2013) who used a variety of machine learning prediction algorithms for the prediction of OC on the same dataset. The study by Stevens *et al.*, (2013) partitioned the data by landuse or organic/mineral soil types, and they obtained best calibrations with a R^2 of 0.76 and 0.78 for mineral and organic soils, respectively.

4.5.2 Multi-tasking prediction

Fig. 4.5 shows the evolution of the error when predicting more properties in a single model. For most properties, we can observe a decrease in the error, ranging from 20.6% for sand content to 74.4% for total N, when we predicted all six properties simultaneously. This behaviour is similar to what Ramsundar *et al.*, (2015) described in their drug discovery study, where accuracy increased continuously with the number of tasks. In the case of pH, there was an increase of the error by 8.3%. Accepting this trade-off will depend on the application. In this case, we think that an increase in the RMSE of 0.04 pH units is acceptable considering the improvements in the predictions of the other properties.

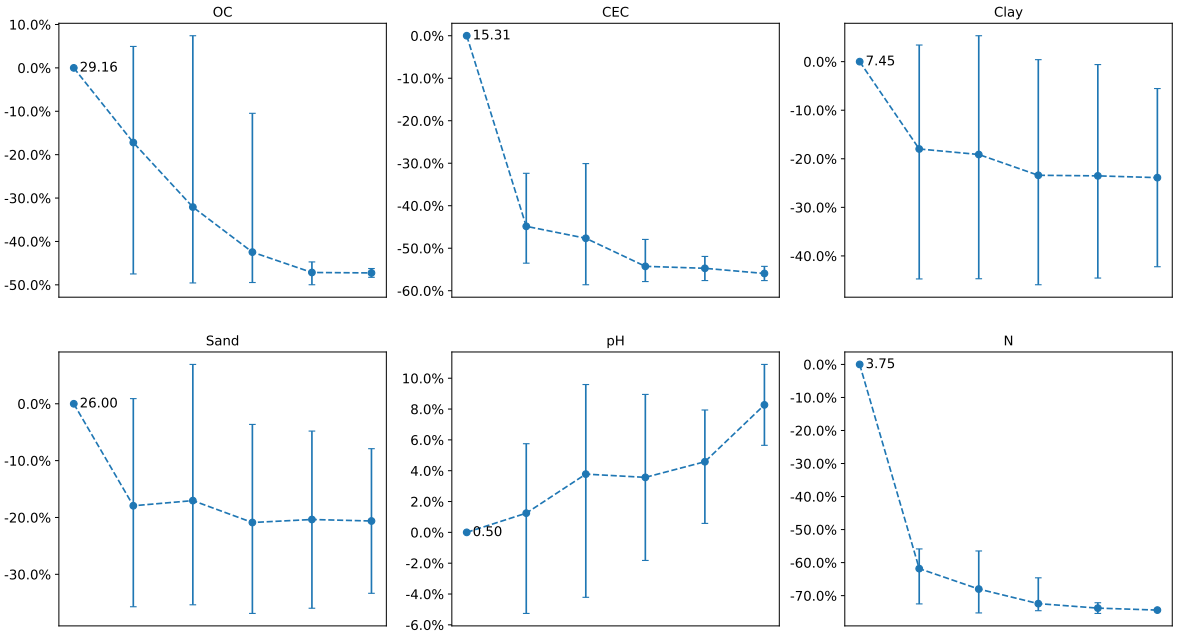


Fig. 4.5: Percentage change in error when more properties are predicted simultaneously. X-axes correspond to the number of extra variables used, starting from zero. Value next to the first point corresponds to the RMSE when only the target property is used. Error bars correspond to the 90% confidence interval after 100 iterations.

This approach provides an interesting way for simultaneous prediction of soil properties within a single model, not just in the sense of using a single spectrum for the characterisation of various soil properties (Islam *et al.*, 2003), but as a truly synergistic model. When the network is predicting a property, it uses the rest of the

predicted properties as "hints" that constrain or enhance the prediction. A simplified example is in the case of clay and sand content. If the model is predicting a very high clay content, there is an indication that the sand content should be low. The interactions between 6 properties are more complex, but the decrease in the prediction error when including more soil properties in the prediction, as shown in Fig. 4.5, is an indicator that the network is generating these interaction rules.

4.5.3 CNN vs conventional prediction techniques

To gauge the performance of our CNN, we compared its performance with techniques commonly used in soil spectroscopy literature (Table 4.6), i.e. PLS regression and Cubist regression tree. A graphical summary of the performance of all models is shown in Fig. 4.6. In general, PLS was the worst performing model. Cubist generally presented an excellent performance in the training dataset, but the training/validation error ratio was higher compared with CNN (Cubist: 1.7; CNN single: 1.1; CNN multi: 1.4) which translates into a poorer performance when dealing with data different from the training dataset. The CNN model, when predicting a single property, tended to generalise better in the training dataset, showing a more similar performance between the training and validation dataset. In the multi-task approach, the CNN performed better than the single prediction (as shown in Fig. 4.5) and usually better than the Cubist model.

Based on the test dataset (Table 4.6), Cubist performed better than PLS except for sand and pH. Notably, Cubist showed a relative improvement (calculated as the RMSE difference over PLS) of 67% for OC. All individual CNN models performed better than PLS and Cubist. Compared with Cubist, the single-prediction CNN model reduced RMSE by 26.5, 24.4, 29.3, 17.9, 26.4, and 34.9% for OC, CEC, clay, sand, pH and total N, respectively. The multi-task CNN further improved over Cubist by 61.6, 42.6, 31.7, 23.0, 22.3, and 55.4% for the same soil properties. The total improvement of the multi-task CNN over PLS was 87.1, 52.1, 16.7, 12.8, 13.1, and 67.0%. Such dramatic improvements are rarely found in the soil spectroscopy literature. For example, in Stevens *et al.*, (2013) study for OC prediction, the maximum difference in RMSE for different machine learning methods is about 20%.

OC is usually the focus of many soil spectroscopy studies (Dalal and Henry,

Table 4.6: Comparison of the performance of all methods for the test dataset for OC (g kg^{-1}), CEC ($cmol^{+}$ kg^{-1}), clay content (%), sand content (%), pH and N (g kg^{-1}).

		PLS	Cubist	CNN	CNN_multi
OC	RMSE	130.50	43.75	32.14	16.82
	R^2	0.35	0.79	0.88	0.69
	ME	-3.97	-2.23	-2.28	2.25
	ρ_c	0.52	0.89	0.94	0.83
CEC	RMSE	13.60	11.34	8.58	6.51
	R^2	0.23	0.41	0.66	0.63
	ME	-1.47	0.06	1.86	-0.87
	ρ_c	0.47	0.64	0.75	0.77
Clay	RMSE	8.75	10.67	7.55	7.29
	R^2	0.55	0.42	0.70	0.68
	ME	-0.53	-0.16	-2.67	-0.18
	ρ_c	0.71	0.64	0.81	0.81
Sand	RMSE	19.49	22.09	18.15	17.00
	R^2	0.44	0.38	0.53	0.59
	ME	-0.45	1.00	-1.58	-0.45
	ρ_c	0.63	0.62	0.70	0.76
pH	RMSE	0.61	0.68	0.50	0.53
	R^2	0.80	0.77	0.87	0.84
	ME	0.02	0.02	0.06	-0.08
	ρ_c	0.89	0.87	0.93	0.91
N	RMSE	3.21	2.37	1.54	1.06
	R^2	0.43	0.64	0.83	0.60
	ME	-0.42	-0.08	-0.34	0.01
	ρ_c	0.65	0.80	0.90	0.77

1986; Ben-Dor and Banin, 1995; Chang *et al.*, 2001; McCarty *et al.*, 2002; Wills *et al.*, 2014) given that it is a key component of functional ecosystems and crucial for food, soil, water and energy security (Stockmann *et al.*, 2015; Minasny *et al.*, 2017). The multi-task CNN is able to substantially improve the prediction compared with traditional methods, over a wide range of soils, which makes it an interesting candidate for rapid soil carbon assessment model in a regional or global context.

It is also worth mentioning that the CNN models were trained with little spectral pre-treatment, we used the full reflectance spectra, and utilised all the multiscale

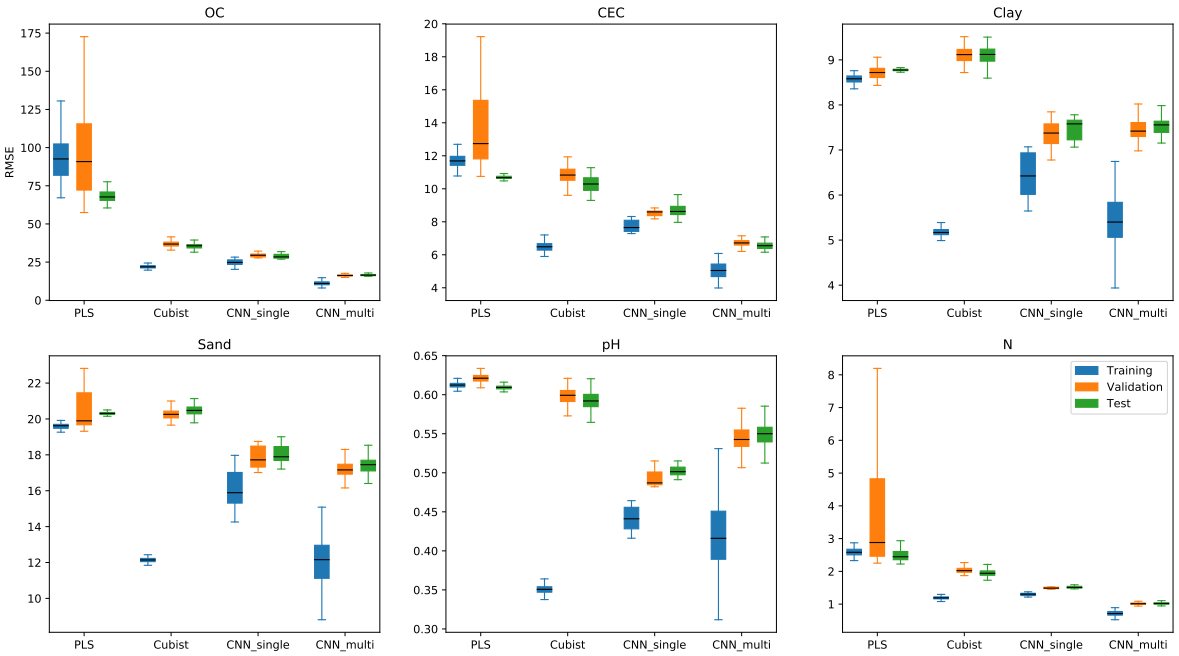


Fig. 4.6: Comparison between PLS, Cubist and CNN for OC (g kg^{-1}), CEC ($cmol^{+}$ kg^{-1}), clay content (%), sand content (%), pH and N (g kg^{-1}). The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR).

variation represented by the spectrogram. We believe that the CNN is able to select the best spectra representation based on various smoothing scales. Previous approaches require decomposition of the spectra using wavelets, and selecting wavelet coefficients to produce parsimonious multivariate calibrations (Viscarra Rossel and Lark, 2009).

CNN is known to outperform conventional prediction techniques and this has been found in different disciplines (LeCun *et al.*, 2015). Our findings also agree with the study of Bjerrum *et al.*, (2017), which compared PLS and CNN models on spectral data in chemometrics.

4.5.4 Dataset size

Machine learning, and especially deep learning, is a very data-hungry approach. Our proposed method has proved to work well with a heterogeneous training dataset of around 10,000 samples (considering that 25% of the initial dataset was excluded as a test set, and approximately 1/3 of data not sampled by the bootstrapping routine

and used as an internal validation set). However, local soil spectroscopy datasets are generally much smaller, and thus we tested our proposed method on a smaller dataset of 390 soil samples.

For all the methods, the R^2 values (Table 4.7) were within an expected range. The RMSE values (Fig. 4.7) showed a good performance of CNN predicting a single property. In comparison with the Cubist model, the trend is similar to the one observed in the LUCAS dataset (Fig. 4.6), with Cubist performing better in the training set and with a greater training/validation error ratio (Cubist: 2.7; CNN single: 1.6; CNN multi: 1.0).

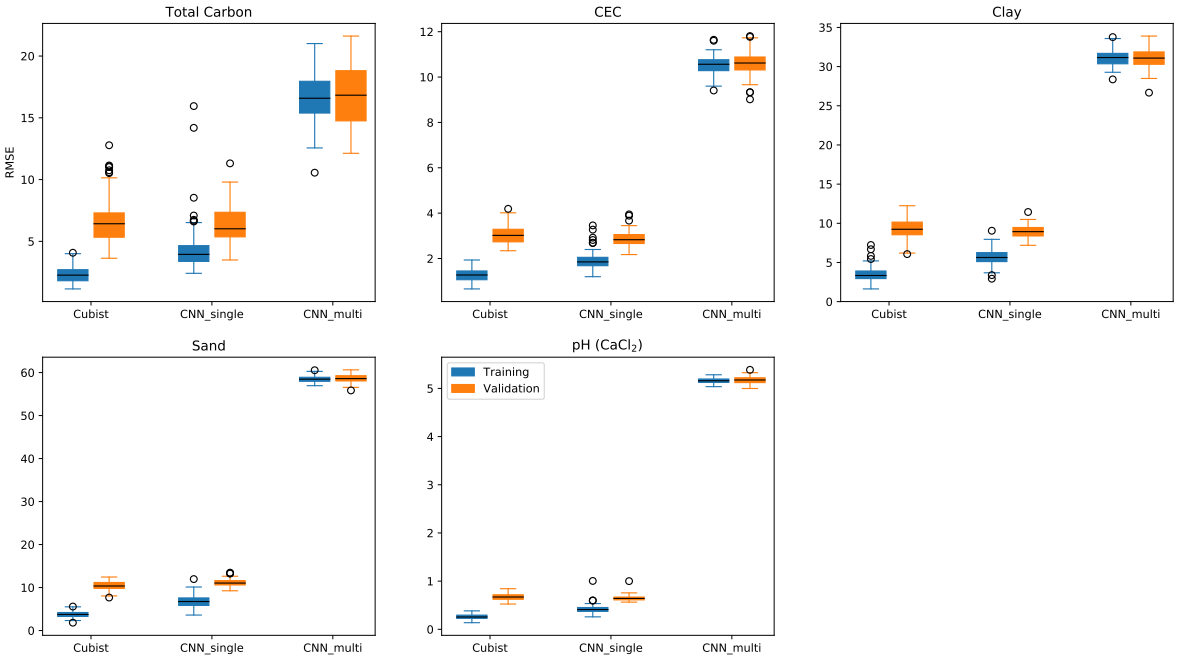


Fig. 4.7: Comparison of error (100 iterations) between training and validation sets for the small dataset. The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR).

We want to draw the attention to the performance of the multi-task CNN. For all properties, the performance is worse than the Cubist model, probably due to the lack of information (small number of samples) when trying to generalise. This agrees with the general consensus in deep learning that dataset size matters. Despite this fact, it is possible that this behaviour will vary for small, more homogeneous datasets.

Table 4.7: Comparison of the coefficient of determination (mean R^2 for 100 iterations) of all methods for the small dataset (validation set).

	Cubist	CNN_single	CNN_multi
Total Carbon	0.79	0.79	0.77
CEC	0.74	0.78	0.77
Clay	0.70	0.73	0.72
Sand	0.62	0.58	0.58
pH ($CaCl_2$)	0.57	0.59	0.58

4.6 Conclusions

We successfully applied a CNN model to predict six soil properties from raw spectral data. The only data pre-treatment used was representing the raw spectral data as a spectrogram, a technique not commonly used in soil spectroscopy and, compared with the traditional pre-treatments, very simple to apply.

We also explored the capacity of CNN to be trained in a multi-task setting. This allowed the simultaneous (with one model) prediction of six soil properties from a single spectrum. This has obvious implications in simplicity and computing time, but also the capacity to achieve synergy, usually improving the predictions compared with a single property prediction. This has an interesting potential, especially for large soil spectroscopy projects where the spectral data usually goes along with laboratory measurements of multiple soil properties.

The proposed approach performs better than the traditional PLS regression and Cubist models, which are the most commonly used models for soil spectroscopy. In our study, the multi-task approach significantly reduced the error compared to the Cubist model, notably by 61.6 and 55.4% for organic carbon and total nitrogen, respectively. Such dramatic improvement is rarely found in spectroscopy studies.

We observed that the multi-task CNN was not effective on a smaller dataset. Our approach showed worse performance than the traditional Cubist model. This confirms that deep learning is a very data-hungry approach. This fact has been widely recognised by the deep learning community, which, up to this point, agrees that dataset size matters.

Deep learning is an exciting discipline that has already changed many fields,

including computer vision, natural language processing, medical image analysis, etc. This paper shows how deep learning has the potential to change the way we model soil spectral data.

4.7 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org.

Barnes, R., Dhanoa, M. S., and Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy* 43 (5): 772–777.

Bellon-Maurel, V. and McBratney, A. (2011). Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils–Critical review and research perspectives. *Soil Biology and Biochemistry* 43 (7): 1398–1410.

Ben-Dor, E and Banin, A (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Science Society of America Journal* 59 (2): 364–372.

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*: pp. 17–36.

Bjerrum, E. J., Glahder, M., and Skov, T. (2017). Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics. *arXiv preprint arXiv:1710.01927*.

Blackman, R. B. and Tukey, J. W. (1958). The measurement of power spectra.

Brown, D. J., Shepherd, K. D., Walsh, M. G., Mays, M. D., and Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* 132 (3): 273–290.

Chang, C.-W., Laird, D. A., Mausbach, M. J., and Hurburgh, C. R. (2001). Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. *Soil Science Society of America Journal* 65 (2): 480–490.

Chen, Y., Lin, Z., Zhao, X., Wang, G., and Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected topics in applied earth observations and remote sensing* 7 (6): 2094–2107.

Chollet, F. *et al.*, (2015). *Keras*. `https://github.com/fchollet/keras`.

Dalal, R. and Henry, R. (1986). Simultaneous Determination of Moisture, Organic Carbon, and Total Nitrogen by Near Infrared Reflectance Spectrophotometry 1. *Soil Science Society of America Journal* 50 (1): 120–123.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Vol. 57. New York: CRC press: p. 436.

Geeves, G., Cresswell, H., Murphy, B., Gessler, P., Chartres, C., Little, I., and Bowman, G. (1995). *The physical, chemical and morphological properties of soils in the wheat-belt of southern New South wales and northern Victoria. NSW Department of Conservation and Land Management*.

Gras, J.-P., Barthès, B. G., Mahaut, B., and Trupin, S. (2014). Best practices for obtaining and processing field visible and near infrared (VNIR) spectra of topsoils. *Geoderma* 214: 126–134.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32 (2): 236–243.

Islam, K., Singh, B., and McBratney, A. (2003). Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Soil Research* 41 (6): 1101–1114.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*: pp. 1097–1105.

Kuhn, M. and Quinlan, R. (2017). *Cubist: Rule- And Instance-Based Regression Modeling*. R package version 0.2.1.

Lawrence, I and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*: 255–268.

LeCun, Y., Bengio, Y., *et al.*, (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361 (10): 1995.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In: *Advances in neural information processing systems*: pp. 396–404.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553): 436–444.

Martens, H. and Naes, T. (1989). *Multivariate calibration*. John Wiley & Sons.

McCarty, G., Reeves, J., Reeves, V., Follett, R., and Kimble, J. (2002). Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Science Society of America Journal* 66 (2): 640–646.

Mevik, B.-H., Wehrens, R., and Liland, K. H. (2016). *pls: Partial Least Squares and Principal Component Regression*. R package version 2.6-0.

Minasny, B. and McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and intelligent laboratory systems* 94 (1): 72–79.

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B. S., *et al.*, (2017). Soil carbon 4 per mille. *Geoderma* 292: 59–86.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., and Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering* 152: 104–116.

Niazi, N. K., Singh, B., and Minasny, B (2015). Mid-infrared spectroscopy and partial least-squares regression to estimate soil arsenic at a highly variable arsenic-contaminated site. *International Journal of Environmental Science and Technology* 12 (6): 1965–1974.

Nitish, S., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15 (1): 1929–1958.

Potter, R., Kopp, G., and Green, H. (1947). Visible Speech. *Van Nostrand, New York*.

Python Software Foundation (2017). *Python Language Reference*. Python Software Foundation.

Quinlan, J. R. *et al.*, (1992). Learning with continuous classes. In: *5th Australian joint conference on artificial intelligence*. Vol. 92. Singapore: pp. 343–348.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*.

Romero, D. J., Ben-Dor, E., Demattê, J. A., e Souza, A. B., Vicente, L. E., Tavares, T. R., Martello, M., Strabeli, T. F., da Barros, P. P. S., Fiorio, P. R., *et al.*, (2018). Internal soil standard method for the Brazilian soil spectral library: Performance and proximate analysis. *Geoderma* 312: 95–103.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sarathjith, M., Das, B. S., Wani, S. P., and Sahrawat, K. L. (2016). Variable indicators for optimum wavelength selection in diffuse reflectance spectroscopy of soils. *Geoderma* 267: 1–9.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* 36 (8): 1627–1639.

Scherer, D., Müller, A., and Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. *Artificial Neural Networks–ICANN 2010*: 92–101.

Shepherd, K. D. and Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil science society of America journal* 66 (3): 988–998.

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., and McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews* 49 (2): 139–186.

Stenberg, B., Rossel, R. A. V., Mouazen, A. M., and Wetterlind, J. (2010). Chapter five-visible and near infrared spectroscopy in soil science. *Advances in agronomy* 107: 163–215.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., and van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PloS one* 8 (6): e66409.

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., Hong, S. Y., Rawlins, B. G., and Field, D. J. (2015). Global soil organic carbon assessment. *Global Food Security* 6: 9–16.

Vašát, R., Kodešová, R., Klement, A., and Borůvka, L. (2017). Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation. *Geoderma* 298: 46–53.

Viscarra Rossel, R. and Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 158 (1): 46–54.

Viscarra Rossel, R. and Lark, R. (2009). Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. *European Journal of Soil Science* 60 (3): 453–464.

Viscarra Rossel, R., Cattle, S. R., Ortega, A., and Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. *Geoderma* 150 (3-4): 253–266.

Viscarra Rossel, R., Behrens, T., Ben-Dor, E., Brown, D., Demattê, J., Shepherd, K., Shi, Z, Stenberg, B, Stevens, A, Adamchuk, V, *et al.*, (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews* 155: 198–230.

Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., and West, L. T. (2014). Overview of the US rapid carbon assessment project: sampling design, initial summary and uncertainty estimates. In: *Soil carbon*. Springer: pp. 95–104.

Chapter 5

Transfer learning to localise a continental soil vis-NIR calibration model

Summary

The rapid development in NIR and information technologies saw the development of various initiatives that have generated large scale databases of soil spectroscopy globally. Models generated within a specific spectral or geographical domain should be carefully used in other contexts since they may lose their validity. This includes the application of global, continental or national spectral libraries to local areas. Both, global and local models are valuable and, ideally, we would like to transfer some of the rules learnt by the more general global models to a local domain. In machine learning, the process of sharing intra-domain information is known as transfer learning. This paper aims to describe and evaluate the effectiveness of transfer learning to "localise" a general soil spectral model. The transfer process consists of, first, training a model with a big volume of data covering a diverse group of cases. Second, some layers of the trained neural network are used to build a local model, which is fine-tuned by using a smaller amount of local data. We demonstrated this method using the LUCAS database, a European dataset, comprising spectral data from 21 countries. For each country, we generated three models: a) Global, with data from all except the country of interest;

b) Local, with data from the country; and c) Transfer, pre-trained as the Global model and fine-tuned with data from the country. The results showed that the Transfer model can lower the error (expressed as RMSE) in 91% of the cases, with a mean reduction of RMSE: 10.5, 11.8, 12.0 and 11.5% for organic carbon, cation exchange capacity, clay content and pH, respectively. This paper demonstrates the usefulness of transfer learning for soil spectroscopy, which will enhance the use of global spectral libraries for local application.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. there are no other authors of the publication according to these criteria;

4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and

5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for this publication is:

Padarian, J., Minasny, B. and McBratney, A.B., 2019. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma, 340, pp.279-288.

Contributors	Statement of contribution
José Padarian *Signature:* José Padarian *Date:* April 2, 2020	Conceptualisation Data analysis Writing
Budiman Minasny	Writing
Alex McBratney	Writing

5.1 Introduction

Soil property prediction using the visible-near infrared (vis-NIR) reflectance spectra has gained a lot of interest in the past decades. Soil spectral libraries as pioneered by Stoner and Biehl (1980) were developed to facilitate the prediction of soil properties from spectral reflectance collected via proximal or remote sensing. The rapid development in NIR and information technologies saw the development of various initiatives that have generated large scale databases of soil spectra globally. Some examples are the RaCA project (Wills *et al.*, 2014; Wijewardane *et al.*, 2016) which collected around 144,000 samples from across the conterminous United States for carbon stock mapping using vis-NIR, or the LUCAS project in Europe with around 20,000 topsoil observations. Viscarra Rossel *et al.*, (2016) compiled a global collection of spectral data of 23,631 observations, mostly focused on Australia, USA and Europe. In addition, national spectral libraries have also been developed to generate country-specific spectral calibration models such as in Denmark (Peng *et al.*, 2013), China (Shi *et al.*, 2014), Brazil (Terra *et al.*, 2015), France (Clairotte *et al.*, 2016) and others.

It is believed that models that were generated within a specific area or soil domain will perform best in that area and perform poorly when applied to other contrasting soil types. In soil science, the high spatial dependency of soil properties means that a model generated for a specific region should be carefully used beyond that spatial domain since they may lose their validity (Minasny *et al.*, 1999; McBratney *et al.*, 2002; Grinand *et al.*, 2008). Conversely, the application of global, continental or national spectral libraries to local areas or regions can also be problematic (Gogé *et al.*, 2014; Ogen *et al.*, 2018). National models often performed poorly when applied in a local area within that country (Wetterlind and Stenberg, 2010). The same is observed when a global model is applied at country level (Mulder *et al.*, 2016). This is understandable, as a global or national model usually captures general trends with a greater generalisation, spanning different soil types. On the other hand, a local area may have short-scale variation which cannot be captured by the global models.

This article used the term “global” model referring to a model calibrated based on a large-scale spectral library, while “local” refers to an area within the “global” dataset. Both, global and local models are valuable and, ideally, we would like to transfer some of the rules learned by the more general global models to a local domain.

In machine learning (ML), this process of sharing intra-domain information is known as transfer learning (Pan and Yang, 2010). This paper aims to describe and evaluate the effectiveness of transfer learning to "localise" a general soil spectral calibration model, simulating a situation where the global dataset is not available for the local user. As far as we know, this is the first time transfer learning has been applied to soil spectroscopy modelling.

5.2 Transfer learning

Humans are capable of applying previously acquired knowledge to tasks that have similar characteristics. Transfer learning, also known as induction learning, is a branch of ML which tries to emulate this process. In our particular case, given a global data domain G and a local data domain L, with $L \subset G$, a traditional ML approach considers both domains as different, generating two independent models, $f(G)$ and $f(L)$ (Fig. 5.1a). In contrast, acknowledging that G and L are somehow related, transfer learning is capable of generating a model $f(L')$ using part of the generalisations learned by $f(G)$ in conjunction with data domain L', with $L' \subseteq L$ (Fig. 5.1b). It is worth noting that, in practice, it is possible that $|L'| \ll |L|$, which gives a considerable advantage to transfer learning, especially when data collection and analysis is limiting.

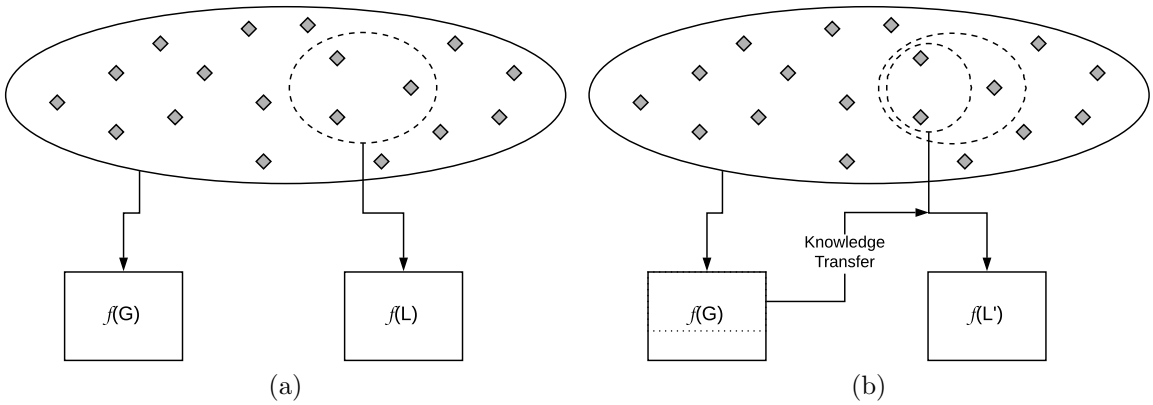


Fig. 5.1: Comparison between (a) traditional machine learning approach and (b) transfer learning.

In this work, after training a global model using the network in Fig. 5.4, the process

consisted on: *a)* extract the first 6 layers (group labelled as "To slice and transfer" in Fig. 5.4), *b)* recreate the same network architecture using as a starting point the 6 trained layers, and *c)* train the new network using local data and keeping the weights of the 6 transferred layers unmodified. A simplified version of the process is also illustrated in Fig. 5.2.

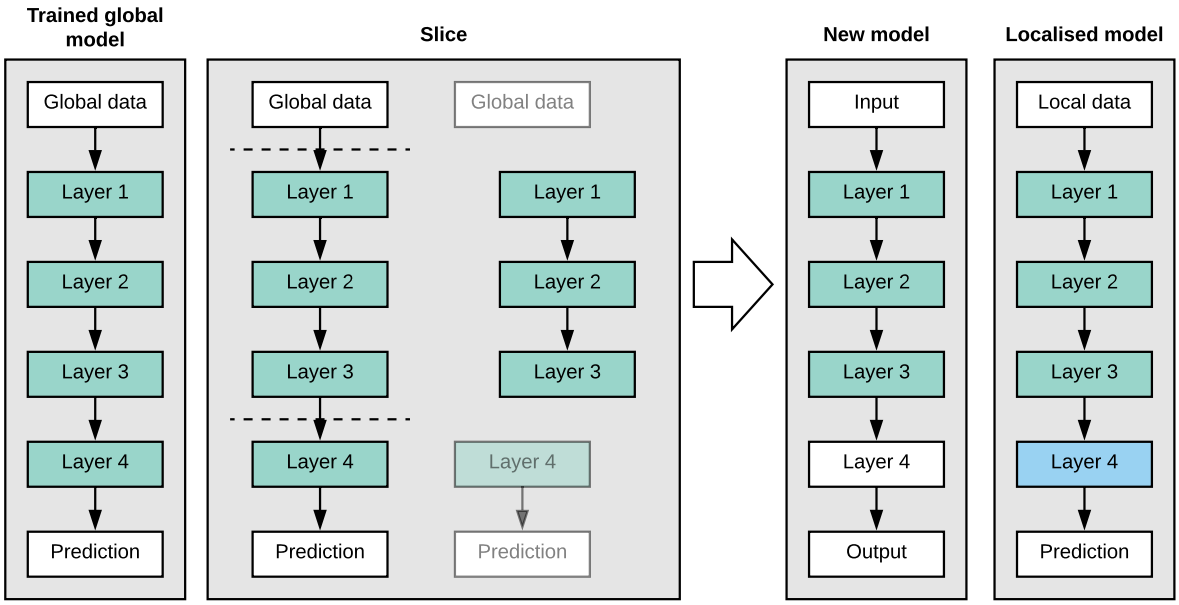


Fig. 5.2: "Localisation" of a global model. Coloured Layers represent trained weight, which are not modified after being learned.

The logic behind this procedure is that in the first training (on the global dataset) the algorithm generates an internal representation of the nature of spectral data. To successfully learn this representation the model needs a big volume of observations, which is exactly what the global dataset provides. In the subsequent training (on the local dataset) the model already "knows" how spectral data behaves, requiring just enough observations to fine-tune the model and adjust it to local conditions.

This approach is commonly used in image recognition (Huh *et al.*, 2016), for example by adjusting a model pre-trained in million of images of different categories and use that knowledge to localise objects in new images (Oquab *et al.*, 2014) or a series of different tasks like image classification, scene recognition, detection of new categories, etc. (Razavian *et al.*, 2014).

5.3 Current approaches

In the soil science literature, two of the most widely used localisation methods are spiking and sub-setting. Spiking consists of adding a small number of local samples to the larger global dataset followed by a re-calibration of the model (Shepherd and Walsh, 2002; Brown, 2007; Wetterlind and Stenberg, 2010). This approach has been reported to show good results (Brown, 2007; Guerrero *et al.*, 2010; Guy *et al.*, 2015), but the method is highly dependent on the size of the global and local datasets (Wetterlind and Stenberg, 2010), and in some cases requires an arbitrary assignment of more weights to the local data (Guerrero *et al.*, 2010). This method biases the predictions towards the local dataset rather than utilising the global knowledge. The results from Gogé *et al.*, (2014) showed that the spiking results are inconsistent. For $CaCO_3$ content, a national model performed best, while for OC and clay content a national model spiked with local samples performed best. Meanwhile for CEC and iron contents local models using 50 local samples were best. These results show that spiking technique is empirical in nature.

Sub-setting consists of selecting a subset of the global dataset that resembles the local data based on some measures of similarity (Araújo *et al.*, 2014), whether spectral similarity or geographical proximity. The major drawback of this technique is that it ignores valuable global information, with the intrinsic risk of generating a less robust model. Similar to spiking, the results depend much on the size of the global and local data. In addition, both approaches assume the availability of global data, which in practice is not always feasible.

5.4 Methods

5.4.1 The data

The LUCAS Soil database consists of around 20,000 topsoil observations with physicochemical and biological properties from Europe (Stevens *et al.*, 2013). The samples were scanned with a diffuse reflectance spectrometer (XDS^{TM} Rapid Content Analyzer), with reflectance data from 400 to 2500nm at a spectral resolution of 0.5nm, resulting in 4200 wavelengths.

For this study we selected four soil properties: *a)* organic carbon content (OC, g kg^{-1}, ISO 10694 1995), *b)* cation exchange capacity (CEC, $cmol^{+}$ kg^{-1}, ISO 11260 1994), *c)* clay particle size fraction (%, ISO 11277 1998), and *d)* pH measured in water (1:5 soil/water ratio, ISO 10390 1994). A summary of statistics of these soil properties is presented in Table 5.1.

Table 5.1: Summary statistics of soil properties for the LUCAS Soil database.

	OC (g/kg^{-1})	CEC ($cmol^{+}$ kg^{-1})	Clay (%)	pH
Minimum	0.00	0.00	0.00	3.21
Maximum	586.80	234.00	79.00	10.08
Mean	50.00	15.76	18.88	6.20
Median	20.80	12.40	17.00	6.21
St. Dev.	91.31	14.48	13.00	1.35
Skewness	3.67	4.24	0.91	-0.08

The data includes information from 21 countries with a variable number of samples for each (Table 5.2).

Table 5.2: Countries in the database used in this study and their corresponding number of samples.

Country	Samples	Country	Samples
France	2798	Austria	412
Spain	2584	Czech Republic	402
Germany	1793	Lithuania	340
Sweden	1725	Latvia	314
Poland	1584	Slovakia	264
Finland	1321	Denmark	218
Italy	1175	Netherlands	198
United Kingdom	830	Estonia	188
Greece	484	Ireland	179
Portugal	476	Slovenia	109
Hungary	434		

5.4.2 Network architecture

Here we present the networks with the best performing combination of hyper-parameters. We used a variation of the Convolutional Neural Network (CNN) multi-task model proposed by Padarian *et al.*, (2019) to simultaneously predict 4 soil properties (Fig. 5.4; Table 5.3) from the spectrogram derived from the raw spectrum.

The CNN is arranged in a series of layers, taking spectrograms as input. There are 3 types of layers: Convolutional Layer, Pooling Layer, and Fully-Connected Layer. The Convolutional layers act as a filter, extracting specific information from the spectrogram, the Pooling layer reduces the spatial resolution of the spectrogram, and a fully-connected layer connects all outputs from previous layers to produce the desired output. We refer the reader to (LeCun *et al.*, 1990) and (Krizhevsky *et al.*, 2012) for the theory behind CNN.

As indicated earlier, rather than a spectrum as input, our CNN takes a spectrogram (Fig. 5.3), a 2D representation of the spectrum showing the reflectance as a function of wavelength and frequency. The spectrogram is obtained by decomposing the signal into overlapping segments using a short-time fast Fourier transformation (Griffin and Lim, 1984). We generated the power spectral density spectrograms by using a Hann window (Blackman and Tukey, 1958) with a span of length 100, with 50% overlap, and a sampling frequency of one. After generating the spectrograms, they were transformed to a logarithmic scale. As a result of this process, we transformed the 1D spectral vector of length 4,200 to a 2D array of shape 51x83 (frequency x wavelength), which is more appropriate for a CNN to process.

This network utilises a series of convolutional layers to analyse the spectrogram, generalising weights for *all* the properties in the "Common layer" section of the network (Table 5.3), to subsequently generate specific weights for *each* property in their respective branch. This type of architecture generates better result than a single-task model because the properties usually share common features in the spectral domain (e.g., peaks, valleys) which a neural network can exploit to better understand the data.

The global and local models were trained on the whole network (Fig. 5.4). For the Transfer model, we used the same network architecture, but preserving the weights of the first 6 layers after being trained with the global data. Subsequently, we trained the Transfer model with local data to adjust the weights of the branches.

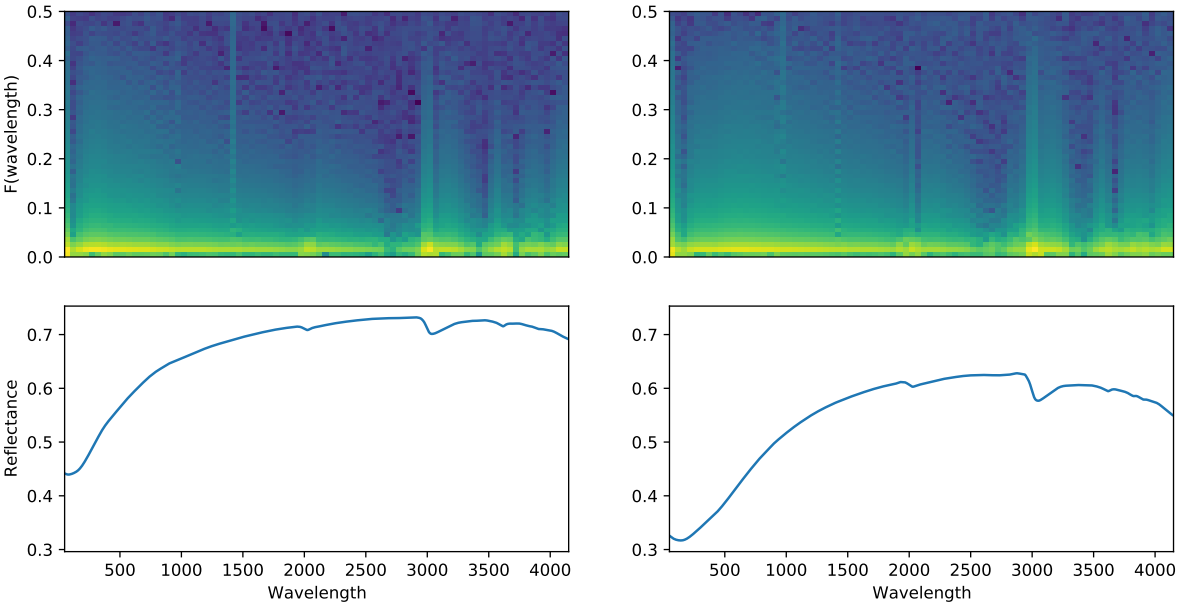


Fig. 5.3: Example of spectral data encoded as a spectrogram. Top panels: spectrogram with amplitude (colour) in log scale. Bottom panels: original spectral data. Left panels: mineral soil (0.5% organic carbon). Right panels: organic soil (20% organic carbon). Reprinted from Padarian *et al.*, (2019).

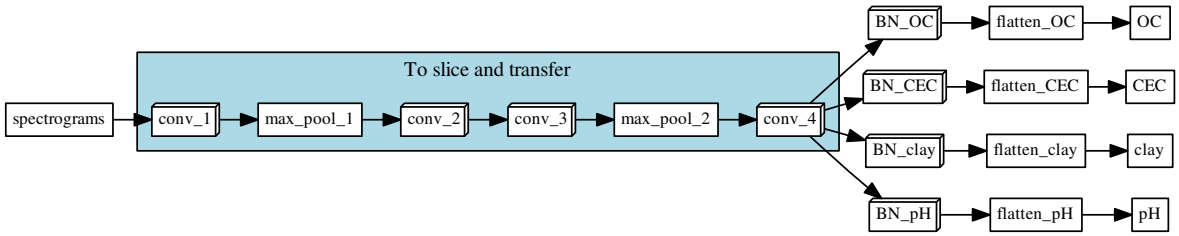


Fig. 5.4: Architecture of the multi-task network used for all the models. Each branch, one per predicted soil property, correspond to a series of one convolutional layer (BN: bottle-neck layer, which reduces the dimensionality of the data) which is then flattened (to 1D) and a fully-connected layer of size=1, which corresponds to the final prediction. The global and local models uses the whole network, as is. The Transfer model uses the "sliced" layers (after being trained with the global data) and then it is trained with local data to adjust the weights of the branches.

5.4.3 Training & Validation

For each country we generated three models:

Local: Only utilising data from the corresponding country. From the available data

Table 5.3: Description of the layers used in this study.

Type	Kernel size	Filters	Activation
†Convolutional	3x3	64	ReLU
†Max-Pooling	2x2	-	-
†Convolutional	3x3	128	ReLU
†Convolutional	3x3	256	ReLU
†Max-Pooling	2x2	-	-
†Convolutional	3x3	512	ReLU
‡Bottle-neck	1x1	64	ReLU
‡Fully-connected	-	1	Linear

†Common layers; ‡for each property (branch)

we randomly held back 10% as a test dataset. From the remaining data, 90% was used for training and 10% for validation and hyper-parameter selection.

Global: From the whole LUCAS dataset, the data from the corresponding country was excluded. From the remaining data, 90% was used for training and 10% for validation and hyper-parameter selection.

Transfer: The training set of the corresponding country was used to localise the Global model previously trained with the global data, as detailed in Section 5.2. The validation set of the corresponding country was used for hyper-parameter selection.

All models were executed with 100 realisations where the training and validation sets were randomly selected. All the models were evaluated (calculating the root mean squared error (RMSE) and coefficient of determination (R^2)) using the test datasets held back from the local model (one per each country). We trained the models during 20 epochs (each epoch is a pass through the whole data). The initial learning rate for the Local and Global model were set to 0.001. For the Transfer model, the initial learning rate was smaller (0.0001) which in our study proved to be a good way to preserve the source model performance. For all the models, we used a batch size of 10 and the *Adam* optimiser (Kingma and Ba, 2014), which maintains a learning rate for each parameter and modifies them during the training based on exponential decays. We compared the models to look for significant differences using a Conover's test (Conover, 1980) with

Bonferroni correction.

5.4.4 Implementation

The CNN was implemented in Python (v3.6.2; Python Software Foundation, 2017) using Keras (v2.1.2; Chollet, 2015) with Tensorflow (v1.4.1; Abadi *et al.*, 2015) backend. Computing was done using the University of Sydney's Artemis high performance computing facility.

5.5 Results and discussion

Figures 5.5–5.7 illustrate the main findings of our study where, for each country, we calculated the RMSE for each of the 3 models and for the 4 soil properties. The plots, arranged from the largest to smallest number of samples, show the distribution of the RMSE for 100 realisations of the validation data. The results show that generally, for countries with larger number of samples (Fig. 5.5), the global model has the largest error followed by the local model, and the most accurate results are achieved using the transfer model. There is no systematic bias of the models (Table A1), with few properties in some countries where there is high bias for the Global model. The bias of the Global model in some countries indicates that distinctive local conditions were not present in the data that were used during the calibration.

Comparing our results with other studies using the LUCAS dataset, our approach generally performs better. The studies by Stevens *et al.*, (2013) and Nocita *et al.*, (2014) report SOC predictions at European scale after training different models for mineral and organic soils, and also grouped by land use. In countries with relatively low SOC levels, like Italy or Spain, our approach yielded similar results (compared with the results reported on mineral soils). In countries with relatively high SOC levels, like Finland or Sweden, our approach performed much better (compared with the results reported on organic soils), reducing the error by around 60%.

5.5.1 Local vs Global models

In general, there is a trend that, in countries with less samples, the global model performed better than the local model (Fig. 5.8 left panel) based on the ratio of performance to inter-quartile distance ($(Q_{75} - Q_{25})/RMSE$). There are a few exceptions for countries with large datasets (i.e., Germany, Czech Republic) but the reasons why the global model performed better are not clear. When we consider geographical sample density (number of samples per unit area of country) (Fig. 5.8 right panel), there is no clear relationship, perhaps indicating that the area of a country does not reflect its pedodiversity.

The number of samples by itself is not a good indicator of the performance of the local model. Considering that the models for each country are evaluated by splitting the national data into a training and a test dataset, the low performance of the local model only implies that it is not possible to consistently draw a representative sample of the national dataset. To properly describe a dataset it is necessary to develop more comprehensive measures that take into account pedodiversity and how much of it is covered by the data.

5.5.2 Transfer learning: Country by country

Looking into more details, for individual cases, in 18 of the 21 countries transfer learning shows a significant improvement for at least one property, compared with the Local and Global models. Four countries showed a significant improvement in all four properties. Even in cases where no significant changes were found, there is a clear trend that the transfer model reduced the mean RMSE. In 14 countries the mean RMSE was reduced for all four properties, and in 76 of the total 84 (90.5%) combinations of countries and properties. The Transfer model yielded a mean RMSE decrease, compared with the second best performing model, of 10.5, 11.8, 12.0 and 11.5% for OC, CEC, Clay and pH, respectively.

In few cases, the Transfer model performed worse than the Global or Local models (such as Finland, United Kingdom, Greece, Denmark, Netherlands and Ireland; Fig. 5.5–5.7), this is when the error of the global model is much larger than the local one, producing negative transfer. In some cases (Italy (Fig. 5.5), Austria (Fig. 5.6)), the Local model was much worse than the Global model and the transfer was not

aggressive enough. In practice, when developing a local model, it is always useful to determine if the source model (global) is beneficial or not, or if a more aggressive transfer is needed, by tracking the performance of the models during training. Ideally, when a positive transfer occurs, the error of the transfer model should show lower initial and final magnitudes as shown in Fig. 5.9.

It is important to note that we used a systematic modelling approach, setting all the models with the same parameters. Considering the sensitivity of the optimisation algorithms to hyper-parameter selection (Bergstra and Bengio, 2012), we believe that it is possible to further fine-tune individual models for each country, and thus increasing the impact of transfer learning.

Our results show that global spectral information is useful when training a local model, improving its performance. This challenges the effectiveness of the spectral or geographical sub-setting method, which is commonly used to improve calibration results when using spectral libraries derived from highly variable, large soil datasets (Araújo *et al.*, 2014; Nocita *et al.*, 2014; Pérez-Fernández and Robertson, 2016). Sub-setting, followed by the development of a local model, limits the model's exposure to broader observations. A good model should be able to utilise data representations with a higher level of abstraction, which are extracted from a wide range of data. This broad data exposure allows a better generalisation and increases the robustness of the model. An opposite conclusion was presented by Guerrero *et al.*, (2016) when using spiking. They concluded that large spectral libraries are less effective for quantitative analysis at local scale. In that case, we think it is mostly a model (partial least squares regression) limitation instead of a problem with the spectral library being "too large".

5.5.3 Effect of number of samples

In general, the prediction accuracy is not dependent on the number of samples within each country, with the relationship between the number of observations and the prediction error not showing a clear trend (Fig. 5.10). The only exception is the case of the Local model for pH, however the results are not conclusive as there are few countries with a large number (> 1000) of samples.

5.5.4 A multi-agent approach

As mentioned in Section 5.2, the required number of samples L' used to localise the global model, in practice, could be smaller than the number of samples L required to generate an independent local model ($|L'| \ll |L|$). This is an important aspect of transfer learning, and modelling in general, which allows *some groups* to take fewer soil samples to obtain good results. We would like to highlight that this implies that *someone* needs to have the responsibility to compile and provide accurate soil information to generate a global model in the first instance. Deep learning is a very powerful approach, but at the same time it is very "data-hungry". We believe that multi-level collaboration is key to feed this virtuous cycle.

5.6 Conclusions

Transfer learning proved to be effective to localise a general soil spectral calibration model generated with a continental dataset. For most of the countries considered in this study, there was an improvement compared with using either a Global (the general model) or Local model (generated only with the data from the respective country).

Our findings also highlight the importance of global databases. They are crucial for understanding processes at the planetary scale but also important to complement our knowledge at a local scale. Collaboration can be beneficial for everyone, even for data-rich countries or organisations.

The transfer learning model does not require the global dataset to be available for local training. Once a global model has been calibrated, only the model that needs to be re-trained can be shared. This is a potential solution for the issue of data privacy. It is important to keep in mind that the proposed method is also applicable if the global dataset is available to a local user.

In this study, we demonstrated the application of transfer learning at a continental scale. It would be useful for future studies to look into the application of the transfer model at a local scale, e.g. transferring a national spectral knowledge to a field.

The concept of transfer learning has proven to be effective in diverse areas, and we envision that it can also be applied to other sub-disciplines of soil science like Digital Soil Mapping.

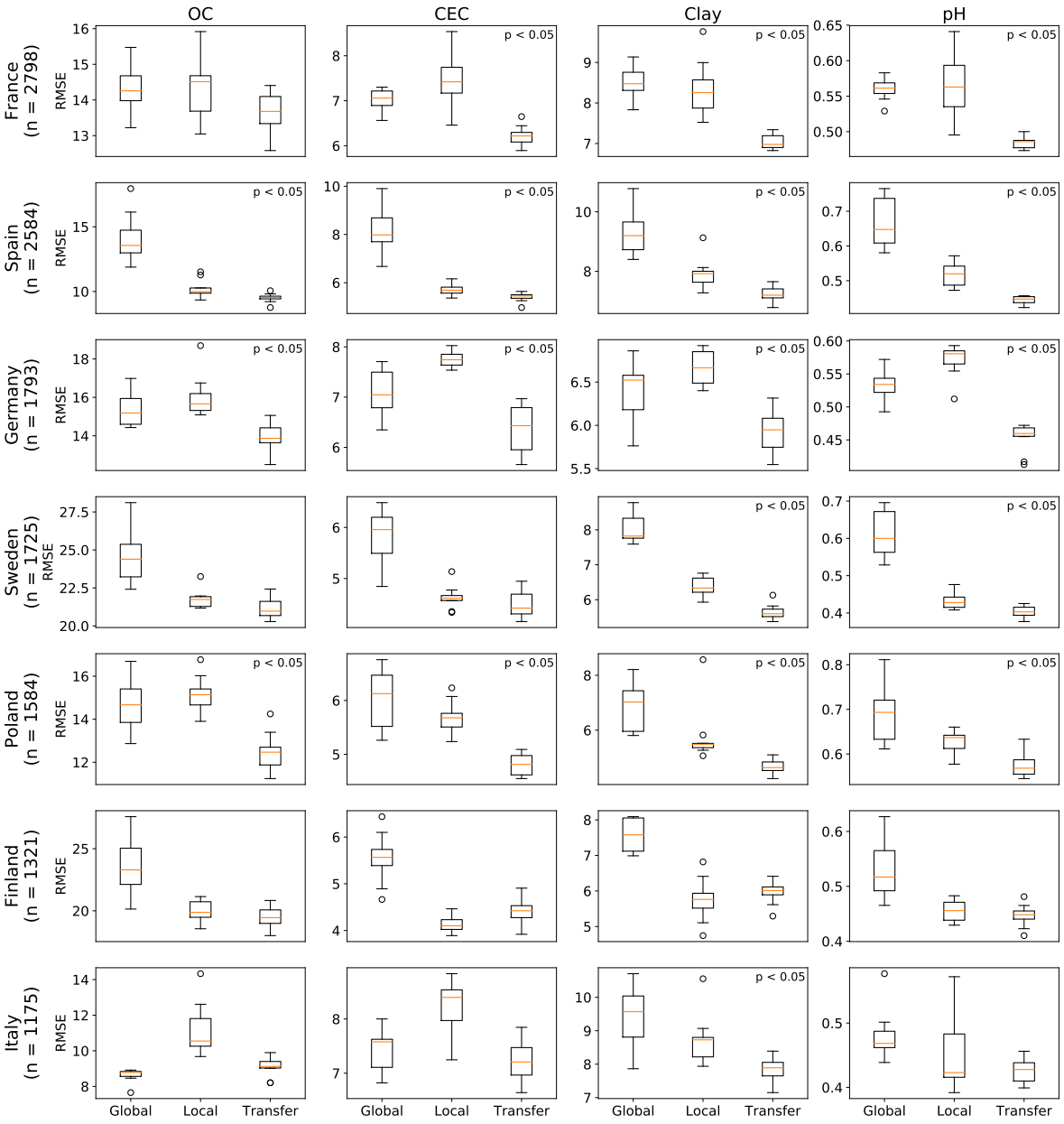


Fig. 5.5: Comparison of RMSE (100 realisations of the validation data) for global, local and transfer models for each country. The error was measured in the test dataset. When the upper-right corner of the panel has a "$p < 0.05$", the transfer model is significative different than both contenders (Conover's test with Bonferroni correction). The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR).

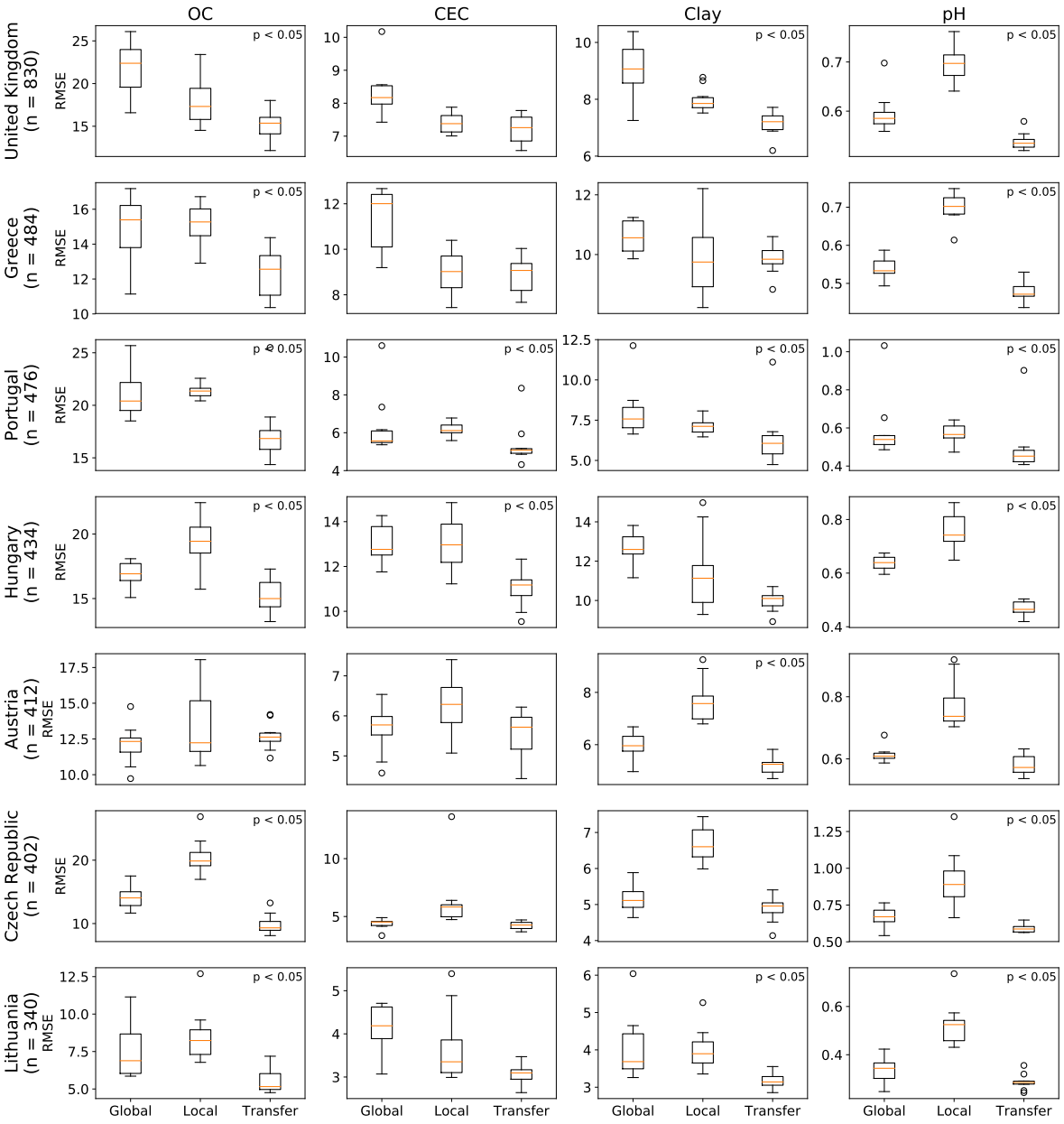


Fig. 5.6: Comparison of RMSE (100 realisations of the validation data) for global, local and transfer models for each country (continuation). The error was measured in the test dataset. When the upper-right corner of the panel has a "$p < 0.05$", the transfer model is significative different than both contenders (Conover's test with Bonferroni correction). The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR).

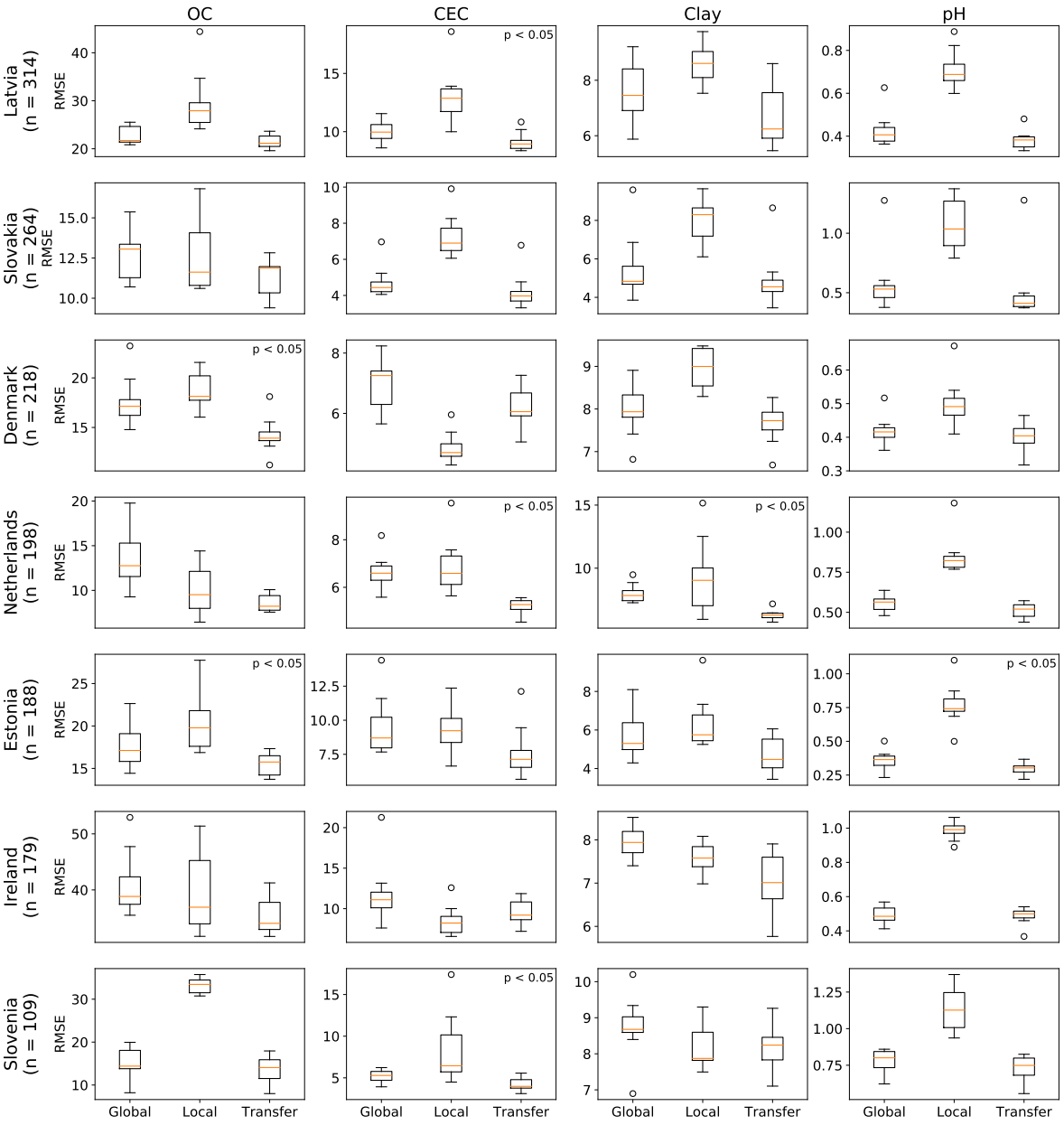


Fig. 5.7: Comparison of RMSE (100 realisations of the validation data) for global, local and transfer models for each country (continuation). The error was measured in the test dataset. When the upper-right corner of the panel has a "$p < 0.05$", the transfer model is significative different than both contenders (Conover's test with Bonferroni correction). The box represents the inter-quartile range (IQR) with a line at the median. Whiskers extends from the first datum greater than (Q1 - 1.5*IQR) to the last datum less than (Q3 + 1.5*IQR).

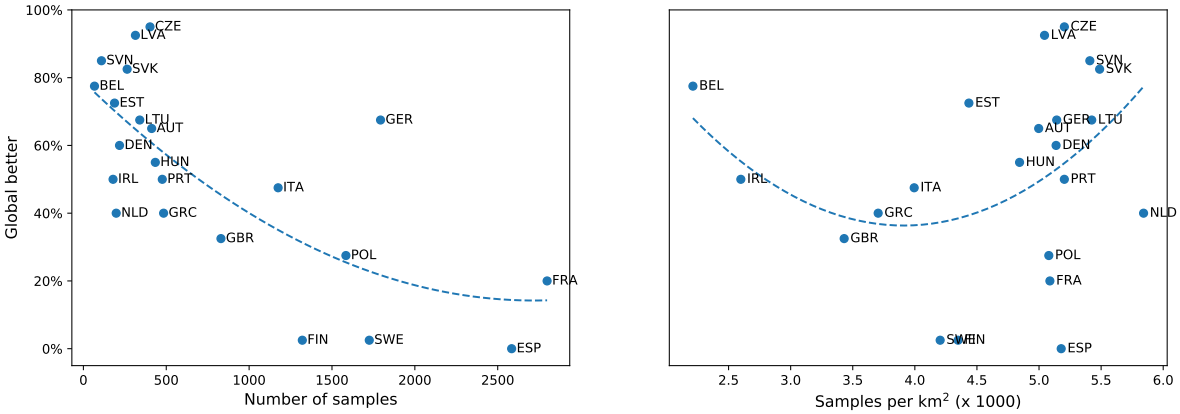


Fig. 5.8: Proportion of repetitions where the Global model performed better than the Local model (based on the ratio of performance to inter-quartile distance).

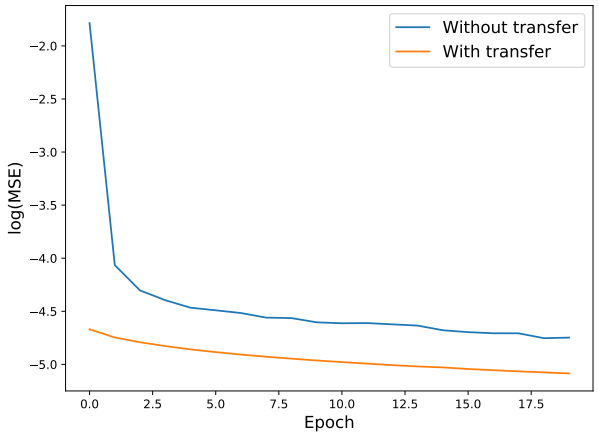


Fig. 5.9: Loss curves for local models with and without transfer. This example is for clay content in France. An epoch corresponds to a pass through all the samples in the dataset.

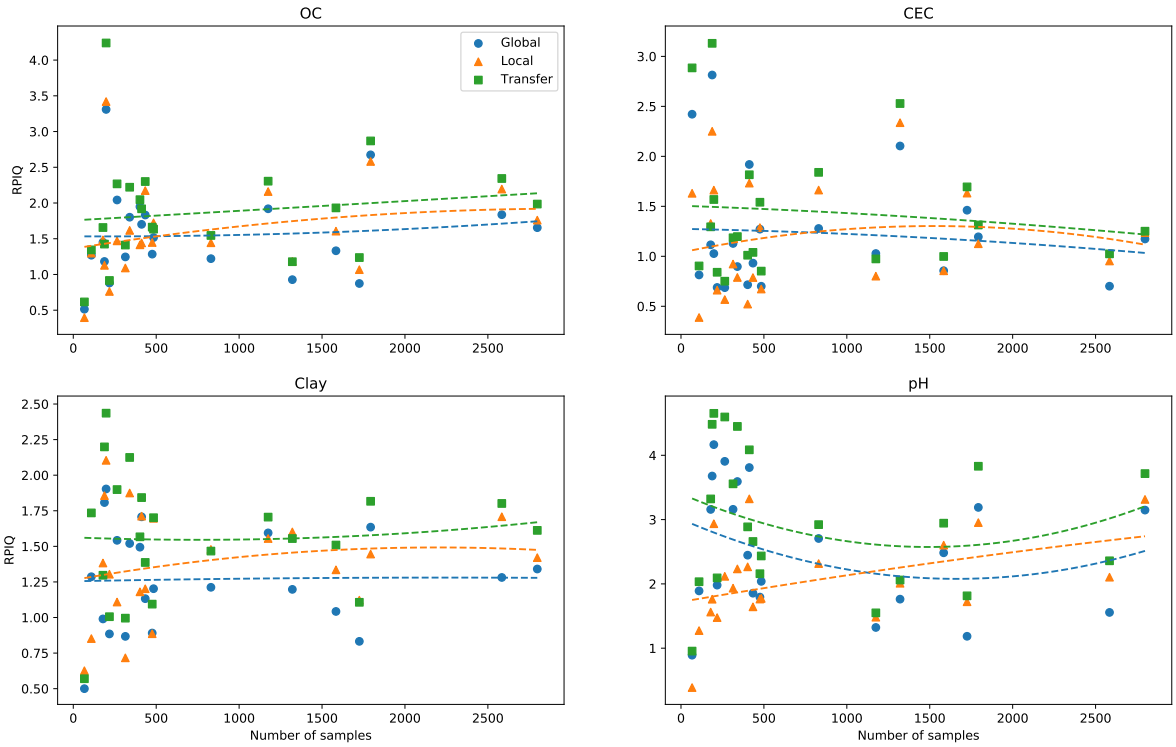


Fig. 5.10: Relationship between RPIQ and number of samples for each property.

5.7 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org.

Araújo, S., Wetterlind, J, Demattê, J., and Stenberg, B (2014). Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *European Journal of Soil Science* 65 (5): 718–729.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13 (Feb): 281–305.

Blackman, R. B. and Tukey, J. W. (1958). The measurement of power spectra.

Brown, D. J. (2007). Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. *Geoderma* 140 (4): 444–453.

Chollet, F. *et al.*, (2015). *Keras*. `https://github.com/fchollet/keras`.

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N. P., Bernoux, M., and Barthès, B. G. (2016). National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. *Geoderma* 276: 41–52.

Conover, W. J. (1980). Practical nonparametric statistics.

Gogé, F., Gomez, C., Jolivet, C., and Joffre, R. (2014). Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? *Geoderma* 213: 1–9.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32 (2): 236–243.

Grinand, C., Arrouays, D., Laroche, B., and Martin, M. P. (2008). Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. *Geoderma* 143 (1-2): 180–190.

Guerrero, C., Zornoza, R., Gómez, I., and Mataix-Beneyto, J. (2010). Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma* 158 (1-2): 66–77.

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A. M., Gabarrón-Galeote, M. A., Ruiz-Sinoga, J. D., Zornoza, R., and Rossel, R. A. V. (2016). Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil and Tillage Research* 155: 501–509.

Guy, A. L., Siciliano, S. D., and Lamb, E. G. (2015). Spiking regional vis-NIR calibration models with local samples to predict soil organic carbon in two High Arctic polar deserts using a vis-NIR probe. *Canadian Journal of Soil Science* 95 (3): 237–249.

Huh, M., Agrawal, P., and Efros, A. A. (2016). What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*: pp. 1097–1105.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In: *Advances in neural information processing systems*: pp. 396–404.

McBratney, A. B., Minasny, B., Cattle, S. R., and Vervoort, R. (2002). From pedotransfer functions to soil inference systems. *Geoderma* 109 (1–2): 41–73.

Minasny, B., McBratney, A., and Bristow, K. (1999). Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma* 93 (3–4): 225–253.

Mulder, V., Lacoste, M, Richer-de Forges, A., Martin, M., and Arrouays, D (2016). National versus global modelling the 3D distribution of soil organic carbon in mainland France. *Geoderma* 263: 16–34.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., and Montanarella, L. (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry* 68: 337–347.

Ogen, Y., Neumann, C., Chabrillat, S., Goldshleger, N., and Dor, E. B. (2018). Evaluating the detection limit of organic matter using point and imaging spectroscopy. *Geoderma* 321: 100–109.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE: pp. 1717–1724.

Padarian, J, Minasny, B, and McBratney, A. (2019). Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional* 16: e00198.

Pan, S. J., Yang, Q., *et al.*, (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22 (10): 1345–1359.

Peng, Y., Knadel, M., Gislum, R., Deng, F., Norgaard, T., de Jonge, L. W., Moldrup, P., and Greve, M. H. (2013). Predicting soil organic carbon at field scale using a national soil spectral library. *Journal of Near Infrared Spectroscopy* 21 (3): 213–222.

Pérez-Fernández, E. and Robertson, A. J. (2016). Global and local calibrations to predict chemical and physical properties of a national spatial dataset of Scottish soils from their near infrared spectra. *Journal of Near Infrared Spectroscopy* 24 (3): 305–316.

Python Software Foundation (2017). *Python Language Reference*. Python Software Foundation.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE: pp. 512–519.

Shepherd, K. D. and Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil science society of America journal* 66 (3): 988–998.

Shi, T., Chen, Y., Liu, Y., and Wu, G. (2014). Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals. *Journal of hazardous materials* 265: 166–176.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., and van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PloS one* 8 (6): e66409.

Stoner, E. R. and Biehl, L (1980). Development of a digital data base for reflectance-related soil information.

Terra, F. S., Demattê, J. A., and Rossel, R. A. V. (2015). Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. *Geoderma* 255: 81–93.

Viscarra Rossel, R., Behrens, T., Ben-Dor, E., Brown, D., Demattê, J., Shepherd, K., Shi, Z, Stenberg, B, Stevens, A, Adamchuk, V, *et al.*, (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews* 155: 198–230.

Wetterlind, J. and Stenberg, B. (2010). Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science* 61 (6): 823–843.

Wijewardane, N. K., Ge, Y., Wills, S., and Loecke, T. (2016). Prediction of soil carbon in the conterminous United States: Visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project. *Soil Science Society of America Journal* 80 (4): 973–982.

Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., and West, L. T. (2014). Overview of the US rapid carbon assessment project: sampling design, initial summary and uncertainty estimates. In: *Soil carbon*. Springer: pp. 95–104.

Chapter 6

Word embeddings for application in geosciences: development, evaluation and examples of soil-related concepts

Summary

A large amount of descriptive information is available in geosciences. This information is usually considered subjective and ill-favoured compared with its numerical counterpart. Considering the advances in natural language processing and machine learning, it is possible to utilise descriptive information and encode it as dense vectors. These word embeddings, which encode information about a word and its linguistic relationships with other words, lay on a multi-dimensional space where angles and distances have a linguistic interpretation. We used 280,764 full-text scientific articles related to geosciences to train a domain-specific language model capable of generating such embeddings. To evaluate the quality of the numerical representations, we performed three intrinsic evaluations, namely: the capacity to generate analogies, term relatedness compared with the opinion of a human subject, and categorisation of different groups of words. Since this is the first attempt to evaluate word embedding for tasks in the geosciences domain, we created a test suite specific for geosciences.

We compared our results with general domain embeddings commonly used in other disciplines. As expected, our domain-specific embeddings (GeoVec) outperformed general domain embeddings in all tasks, with an overall performance improvement of 107.9%. We also presented an example where we successfully emulated part of a taxonomic analysis of soil profiles which was originally applied to soil numerical data, which would not be possible without the use of embeddings. The resulting embedding and test suite will be made available for other researchers to use and expand.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for this publication is:

Padarian, J. and Fuentes, I., 2019. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. Soil, 5(1), pp.177-187.

Contributors	Statement of contribution
José Padarian *Signature:* José Padarian *Date:* April 2, 2020	Conceptualisation Data analysis Writing
Ignacio Fuentes	Data analysis

6.1 Introduction

Machine learning (ML) methods have been used in many fields of geosciences (Lary *et al.*, 2016) to perform tasks such as classification of satellite imagery (Maxwell *et al.*, 2018), soil mapping (McBratney *et al.*, 2003), mineral prospecting (Caté *et al.*, 2017), and flood prediction (Mosavi *et al.*, 2018). Thanks to their capability to deal with complex nonlinearities present in the data, ML usually outperforms more traditional methods in terms of predictive power (see Chapter 1). The application of ML in geosciences usually prioritise numerical or categorical data over qualitative descriptions, which are usually considered subjective in nature (McBratney and Odeh, 1997). However, it must be taken into account the resources that have been invested in collecting large amounts of descriptive information from pedological, geological and other fields of geosciences. Neglecting descriptive data due to its inconsistency seems wasteful, yet natural language processing (NLP) techniques, which involve the manipulation and analysis of language (Jain *et al.*, 2018), have rarely been applied in geosciences.

For soil sciences, NLP opens the possibility to use a broad range of new analyses. Some examples include general, discipline-wide methods such as automated content analysis (Nunez-Mir *et al.*, 2016) or recommendation systems (Wang and Blei, 2011) which can take advantage of the current literature. More specific cases could take advantage of big archives of descriptive data, like the ones reported by Arrouays *et al.*, (2017). The authors mention examples such as the Netherlands with more than 327,000 auger descriptions covering agricultural, forest and natural lands, or the north-central US with 47,364 pedon descriptions covering 8 states.

Approaches to deal with descriptive data include the work of Fonseca *et al.*, (2002) who proposed the use of ontologies to integrate geographic information of different kinds. At the University of Colorado, Chris Jenkins created a structured vocabulary for geomaterials (`http://instaar.colorado.edu/~jenkinsc/dbseabed/resources/geomaterials/`) using lexical extraction (Miller, 1995), names decomposition (Peckham, 2014) and distributional semantics (Baroni *et al.*, 2012) in order to characterise word terms for use in NLP and other applications. A different approach, perhaps closer to the preferred quantitative methods, is the use of dense word embeddings (vectors) which encode information about a word and its linguistic

relationships with other words, positioning it on a multi-dimensional space. The latter is the focus of this study.

There are many general-purpose word embeddings trained on large corpora from social media or knowledge organisation archives such as Wikipedia (Pennington *et al.*, 2014; Bojanowski *et al.*, 2016). These embeddings have been proven to be useful in many tasks such as machine translation (Mikolov *et al.*, 2013b), video description (Venugopalan *et al.*, 2016), document summarisation (Goldstein *et al.*, 2000), and spell checking (Pande, 2017). However, for field-specific tasks, many researchers agree that word embeddings trained on specialised corpora can capture the semantics of terms better than those trained on general corpora (Jiang *et al.*, 2015; Pakhomov *et al.*, 2016; Roy *et al.*, 2017; Nooralahzadeh *et al.*, 2018; Wang *et al.*, 2018).

As far as we are aware, this is the first attempt to develop and evaluate word embedding for the geosciences domain. This paper is structured as follows: first, we define what word embeddings are, explaining how they work and showing examples to help the reader understand some of their properties. Second, we describe the text data used and the pre-processes required to train a language model and generate these word embeddings (GeoVec). Third, we illustrate how a natural language model can be quantitatively evaluated and we present the first test dataset for the evaluation of word embeddings specifically developed for the geosciences domain. Fourth, we present results of an intrinsic evaluation of our language model using our test dataset and we explore some of the characteristics of the multi-dimensional space and the linguistic relationships captured by the model through examples of soil-related concepts. Finally, we present a simple, illustrative example of how the embedding can be used in a downstream task.

6.2 Word embeddings

Word embeddings have been commonly used in many scientific disciplines, thanks to their application in statistics. For example, one-hot encodings (Fig. 6.1), also known as "dummy variables", have been used in regression analysis since at least 1957 (Suits, 1957). In one-hot encoding, each word is represented by a vector of length equal to the number of classes or words, where each dimension represents a feature. The problem with this representation is that the resulting array is sparse (mostly zeros) and very

large when using large corpora, and also presents the problem of poor estimation of the parameters of the less-common words (Turian *et al.*, 2010). A solution for these problems is the use of unsupervised learning to induce dense, low-dimensional embeddings (Bengio, 2008). The resulting embeddings lay on a multi-dimensional space where angles and distances have a linguistic interpretation.

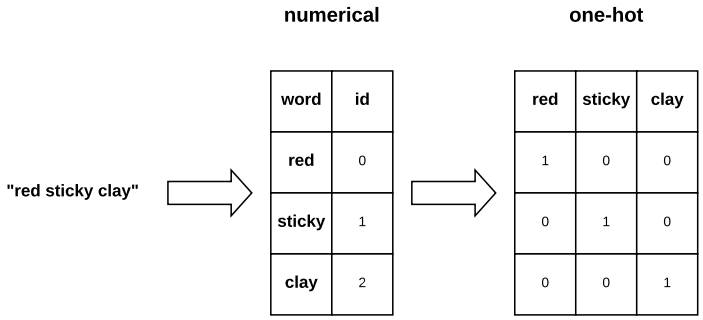


Fig. 6.1: Example of two encodings of the phrase "red sticky clay", numerical and one-hot.

These dense, real vectors allow models, specially neural networks, to generalise to new combinations of features beyond those seen during training thanks to the properties of the vector space where semantically related words are usually close to each other (LeCun *et al.*, 2015). Since the generated vector space also has properties such as addition and subtraction, Mikolov *et al.*, (2013a) gives some examples of calculations that can be performed using word embedding. For instance the operation *vec("Berlin")* - *vec("Germany")* + *vec("France")* generates a new vector. When they calculated the distance from that resulting vector to all the words from the model vocabulary, the closest one was the word "Paris". Fig. 6.2 presents a principal component analysis (PCA) projection of pairs of words with the country-capital relationship. Without explicitly enforcing this relationship when creating the language model, the resulting word embeddings encode the country-capital relationship due to the high co-occurrence of the terms. In Fig. 6.2 it is also possible to observe a second relationship, geographic location, where South American countries are positioned to the right, European countries in the middle and (Eur-)Asian countries to the left.

Potentially, each dimension and interaction within the generated vector space encodes a different type of relationship extracted from the data. Thanks to the properties of the generated vector space, we give ML algorithms the capacity to utilise

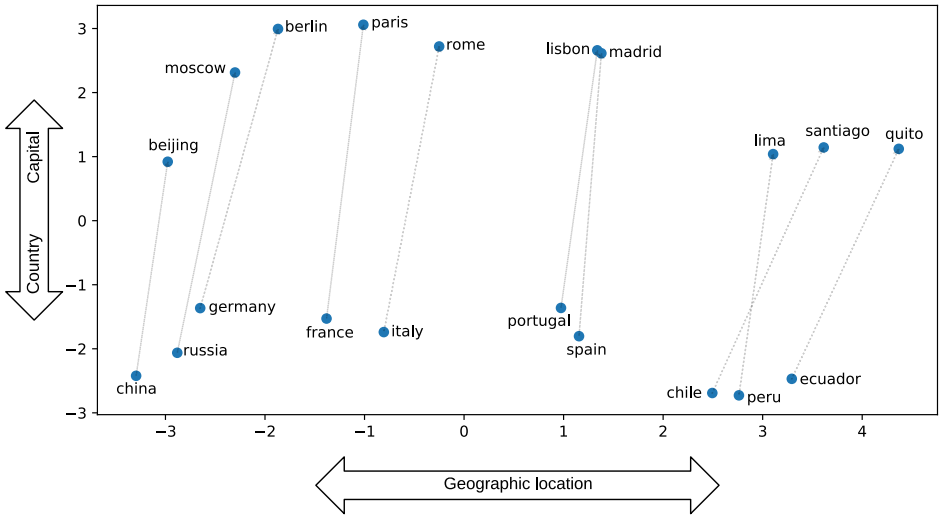


Fig. 6.2: Examples of two-dimensional PCA projection of selected word embeddings using a general domain model. The figure illustrates the country-capital relationship learned by the model. Also notice that the model learned about the geographic relationship between the places. Example adapted from Mikolov *et al.*, (2013a).

and understand text and we are able to use the same methods usually designed for numerical data (e.g. clustering, principal component). In the next sections, we describe how we generated a language model that yields word embeddings that encode semantic and syntactic relations specific for the geosciences domain, we visualise some of those relations and we illustrate how to evaluate them numerically.

6.3 Data, text pre-processing and model training

6.3.1 Corpus

The corpus was generated by retrieving and processing 280,764 full-text articles related to geosciences. We used the Elsevier ScienceDirect APIs to search for manuscripts that matched the terms listed in Table 6.1, which cover a broad range of topics. These terms were selected based on their general relationship with geosciences and specifically soil science. We also included Wikipedia articles which list and concisely define some concepts like types of rocks, minerals, and soils, providing more context than a scientific publication, considering that the model depends on words

co-occurrences. We downloaded the text from Wikipedia articles "List_of_rock_types", "List_of_minerals", "List_of_landforms", "Rock_(geology)", "USDA_soil_taxonomy" and "FAO_soil_classification", and also all the Wikipedia articles linked from those pages.

Table 6.1: Search terms used to retrieve full-text articles from Elsevier ScienceDirect APIs.

Search terms		
Acrisol	Geosciences	Permafrost
Alfisol	Groundwater	Petrology
Allophane	Gypsisols	Podzols
Andisol	Histosol	Sedimentary
Andosols	Hydrogeology	Sedimentary mineralogy
Aridisol	Igneous petrology	Sedimentary petrology
Chernozems	Imogolite	Sedimentary rocks
Entisol	Inceptisol	Sedimentology
Environmental geology	Lithology	Soil classification
Field geology	Metamorphic petrology	Spodosol
Gelisol	Mineralogy	Stratigraphy
Geochemistry	Mollisol	Ultisol
Geology	Oxisol	Vertisol
Geomaterials	Peatland	Volcanic soils
Geomorphology	Pedogenesis	
Geophysics	Pedology	

6.3.2 Pre-processing

The corpus was split into sentences which were then pre-processed using a sequence of commonly used procedures including: *a)* removing punctuation, *b)* lower-casing, *c)* removing digits and symbols, and *d)* removing (easily identifiable) references. The cleaned sentences were then tokenised (split into words). In order to decrease the complexity of the vocabulary, we lemmatised all nouns to their singular form and removed all the words with less than 3 characters. We also removed common English words such as 'the', 'an' and 'most' since they are not discriminating and unnecessarily increase the model size and processing time (a full list of the removed "stop words"

can be found in the documentation of the nltk python library (Bird and Loper, 2004)). Finally, we excluded sentences with less than 3 words. The final corpus has a vocabulary size of 701,415 (unique) words and 305,290,867 tokens.

6.3.3 Model training

For this work, we used the GloVe (Global Vectors) model (Pennington *et al.*, 2014), developed by Stanford University NLP group, which achieved great accuracy on word analogy tasks and outperformed other word embedding models on similarity and entity recognition tasks. As many NLP methods, GloVe relays on ratios of word-word co-occurrence probabilities in the corpus. To calculate the co-occurrence probabilities, GloVe uses a local context window, where a pair of words d words apart contributes to a $1/d$ to the total count. After the co-occurrence matrix X is calculated, GloVe minimises the least-squares objective function

$$\sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T \hat{w}_j + b_i + \hat{b}_j - log X_{ij} \right)^2 \tag{6.1}$$

where X_{ij} is the co-occurrence between the target words i and the context word j, V is the vocabulary size, w_i is the word embedding, $\hat{w}_j$ is a context word embedding, b_i and $\hat{b}_j$ are biases for w_i and $\hat{w}_j$, respectively, and $f(X_{ij})$ is the weighting function

$$f(X_{ij}) = \begin{cases} (X_{ij}/x_{\max})^\alpha & \text{if } X_{ij} < x_{\max} \\ 1 & \text{otherwise} \end{cases} \tag{6.2}$$

that assures that rare and frequent co-occurrences are not overweighted. Pennington *et al.*, (2014) recommend using the values 0.75 for the smoothing parameter α and 100 for the maximum cutoff count $x_{\max}$.

We trained the model during 60 epochs, where 1 epoch corresponds to a complete pass through the training dataset. During the training phase, we experimented using embedding with different number of components (dimensions) and different context window sizes. Here we present the results for 300 components and a context window of size 10, which represents a good balance between model size, training time and performance.

6.4 Evaluation of word embeddings

Given the characteristic of the vector space, the most common method to evaluate word embeddings is to asses their performance in tasks that test if semantic and syntactic rules are properly encoded. Many studies have presented datasets to perform this task. Rubenstein and Goodenough (1965) presented a set of 65 noun synonyms to test the relationship between the semantic similarity existing between a pair of words and the degree to which their contexts are similar. More recent and larger test datasets and task types have been proposed (Finkelstein *et al.*, 2002; Mikolov *et al.*, 2013c; Baroni *et al.*, 2014) but they all have been designed to test general domain vectors. Because this work aims to generate embeddings for the geosciences domain, we developed a test suite to evaluate their intrinsic quality in different tasks, which are described below.

Analogy: Given two related pairs of words, a:b and x:y, the aim of the task is to answer the question "a is to x as b is to?". The set includes 50 quartets of words with different levels of complexity, from simple semantic relationships to more advance syntactic relations. In practice, it is possible to find y by calculating the cosine similarity between the differences of the paired vectors:

$$\frac{(v_b - v_a) \cdot (v_y - v_x)}{\|v_b - v_a\| \|v_y - v_x\|} \tag{6.3}$$

In this case, v_y is the embedding for each word of the vocabulary and y is the word with the highest cosine similarity. Some examples of analogies are: "moraine is to glacial as terrace is to ___? (fluvial)", "limestone is to sedimentary as tuff is to ___? (volcanic)" and "chalcantite is to blue as malachite is to ___? (green)".

We estimated the top-1, top-3, top-5 and top-10 accuracy score, recording a positive result if y was within the first 1, 3, 5 or 10 words returned by the model, respectively.

Relatedness: For a given pair of words (a, b), a score of 0 or 1 is assigned by a human subject if the words are unrelated or related, respectively. The set includes 100 pairs of scored pairs of words. The scores are expected to have a high correlation with the cosine similarity between the embeddings of each pair of words. In this

work, we used the Pearson correlation coefficient to evaluate the model against annotations made by 3 people with a geosciences background.

Categorisation: Given 2 sets of words $s_1 = \{a, b, c, ...\}$ and $s_2 = \{x, y, z, ...\}$, this test should be able to correctly assign each word to its corresponding group using a clustering algorithm. We provide 30 tests with 2 clusters each. We estimated the v-measure score (Rosenberg and Hirschberg, 2007), which takes into account the homogeneity and completeness of the clusters, after projecting the multi-dimensional vector space to a two-dimensional PCA space and performing a k-means clustering. Given that k-means is not deterministic (when using random centroids initiation), we used the mean v-measure score of 50 realisations.

We compared our results with general domain vectors trained on Wikipedia articles (until 2014) and the Gigaword v5 catalogue, which comprise 6 billion tokens and is provided by the authors of GloVe at `https://nlp.stanford.edu/projects/glove/`.

6.5 Illustrative example

In order to illustrate the use of word embedding in a downstream application, we decided to emulate part of the analysis of a soil taxonomic system performed by Hughes *et al.*, (2017). They used 23 soil variables (e.g. sand content and bulk density), in their majority numerical and continuous except for two binary variables representing the presence or absence of water or ice. Those variables correspond to the representation of horizons from soil profiles, which were then aggregated (mean) at different taxonomic levels to obtain class centroids.

Our analysis was similar, but, instead of using soil variables, we used the word embedding corresponding to the textual description of 10,000 soil profile descriptions downloaded from the USDA-NRCS Web Site for Official Soil Series Descriptions and Series Classification. The descriptions were pre-processed using the same pipeline used for the corpus (Section 6.3.2). After obtaining the embeddings for each token in the descriptions, we calculated the mean embedding value per profile, which can be considered as an embedding at the profile level. The profiles and their corresponding 300-dimensional embeddings were aggregated at Great Group (GG)

level (Soil Taxonomy) and a mean embedding value was estimated (equivalent to the centroids obtained by Hughes *et al.*, (2017)). After projecting the GG embeddings into a 2-dimensional PCA space, we computed the convex-hull per soil order (smaller convex polygon needed to contain all the GG points for a particular soil order) as a way of visualising their extent.

6.6 Results and discussion

6.6.1 Co-occurrence

Before training the language model, the first output of the process is a co-occurrence matrix. This matrix encodes useful information about the underlying corpus (Heimerl and Gleicher, 2018). Fig. 6.3 shows the co-occurrence probabilities of soil taxonomic orders and some selected words. It is possible to observe that concepts generally associated with a specific order co-occur in the corpus, such as soil cracks, which are features usually present in Vertisols; or Andisols being closely related to areas with volcanic activity.

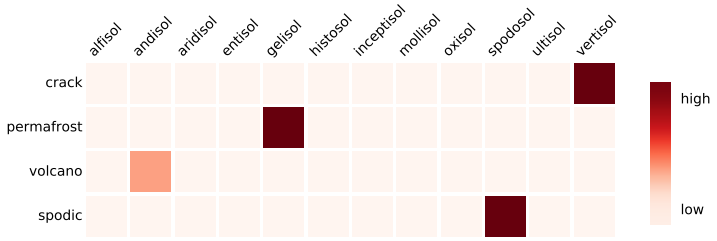


Fig. 6.3: Co-occurrence probability matrix of soil orders (USDA) and selected words.

This information can also be used to guide the process of generating a domain-specific model. In our case, in an early stage of this study, the terms "permafrost" and "gelisol" presented a very low co-occurrence probability, a clear sign of the limited topic coverage of the articles at that point.

6.6.2 Intrinsic evaluation

The results of the intrinsic evaluation indicate that our domain-specific embeddings (GeoVec) performed better than the general domain embeddings in all tasks (Table 6.2),

increasing the overall performance by 107.9%. This is an expected outcome considering the specificity of the tasks. For the analogies, we decided to present the top-1, 3, 5 and 10 accuracy scores because, even if the most desirable result is to have the expected word as the first output from the model, in many cases the first few words are closely related or they are synonyms. For instance, for the analogy "fan is to fluvial as estuary is to ___? (coastal)", the first four alternatives are "tidal", "river", "estuarine", "coastal", which are all related to a estuary.

In the relatedness task, the 3 human annotators had a high inter-annotator agreement (multi-kappa=98.66%; as per Davies and Fleiss (1982)), which was expected since the relations are not complex for some with a background in geosciences. As we keep working on this topic, we plan to extend the test suite with more subtle relations.

Table 6.2: Evaluation scores for each task for our domain-specific (GeoVec) and general domain embeddings (Stanford). For the analogy task, top-1, 3, 5 and 10 represents the accuracy if the expected word was within the first 1, 3, 5 or 10 words returned by the model. For the relatedness task, the score represents the absolute value of the Pearson correlation (mean of the 3 human subjects). For the categorisation task, the score represents the mean value of 50 v-measure scores. The possible range of all scores is 0 to 1, where higher is better.

	GeoVec	Stanford
Analogy (top-1)	0.39	0.22
Analogy (top-3)	0.78	0.37
Analogy (top-5)	0.90	0.41
Analogy (top-10)	0.92	0.49
Relatedness	0.61	0.23
Categorisation	0.75	0.38
Overall	0.73	0.35

It was possible to observe an increase on the overall performance of the embeddings (calculated as the mean of the analogy (top-5), relatedness and categorisation tasks) as we added more articles, almost stabilising around 300 million tokens, especially for the analogy task (Fig. 6.4). For domain-specific embeddings, this limit most likely varies depending on the task and domain. For instance, Pedersen *et al.*, (2007), measuring semantic similarity and relatedness in the biomedical domain, found a limit of around 66 million tokens.

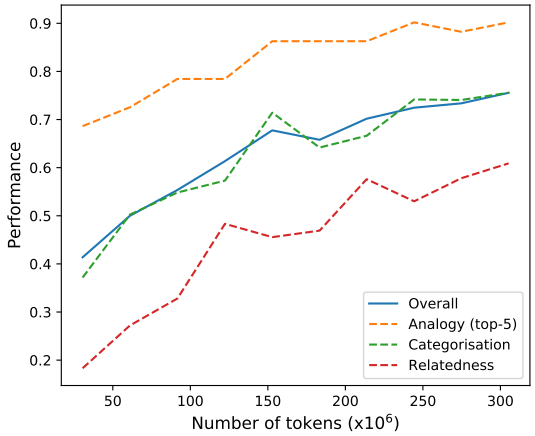


Fig. 6.4: Overall performance of the embeddings versus number of tokens used to construct the co-occurrence matrix. The improvement limit is around 300 million tokens. For future comparisons, this limit corresponds to approximately: 280,000 articles, 22.5 million sentences and 700,000 unique tokens.

The improvement over the general domain embeddings has also been reported in other studies. Wang *et al.*, (2018) concluded that word embeddings trained on biomedical corpora can capture the semantics of medical terms better than the embeddings of a general domain GloVe model. Also in a biomedical application, Jiang *et al.*, (2015) and Pakhomov *et al.*, (2016) reported similar conclusions. In the following sections, we explore the characteristics of the obtained embeddings, showing some graphical examples of selected evaluation tasks.

6.6.3 Analogy

A different way of evaluating analogies is to plot the different pairs of words in a 2-dimensional PCA projection. Fig. 6.5 shows different pairs of words which can be seen as group analogies. From the plot, any pair of related words can be expressed as an analogy. For example, from the left panel, it is possible to generate the analogy "claystone is to clay as sandstone is to ___? (sand)" and the first model output is indeed "sand".

As we showed in Fig. 6.2, the embeddings encode different relationships with different degrees of sophistication. In the left panel of Fig. 6.5 it is possible to observe simple analogies, mostly syntactic since "claystone" contains the word "clay". The

right panel presents a more advanced relationship where rock names are assigned to their corresponding rock type.

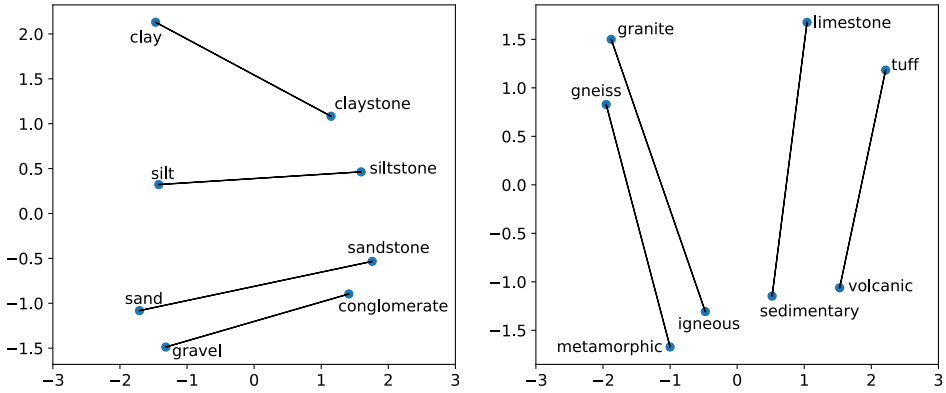


Fig. 6.5: Two-dimensional PCA projection of selected words. Simple syntactic relationship between particle fraction sizes and rocks (left panel) and advanced semantic relationship between rocks and rock types (right panel).

6.6.4 Categorisation

Similar to the analogies, the categorisation task can also present different degrees of complexity of the representations. In the left panel of Fig. 6.6, k-means clustering can distinguish the two expected clusters of concepts, WRB (FAO, 1988) and Soil Taxonomy (USDA, 2010) soil classification names. Andisols and Andosols are correctly assigned to their corresponding groups but apart from the rest, probably due to their unique characteristics. Vertisols are correctly placed in between the two groups since both have a soil type with that name. A second level of aggregation can be observed in the right panel. The k-means clustering correctly assigned the same soil groups from the left panel into a general "soil types" group, different from "rocks".

6.6.5 Other embedding properties

Interpolation of embeddings is an interesting exercise that allows to further explore if the corpus is well represented by the vector space. Interpolation has been used to generate gradient between faces (Yeh *et al.*, 2016; Upchurch *et al.*, 2017), assist drawing (Baxter and ichi Anjyo, 2006) and transform speech (Hsu *et al.*, 2017). Interpolation

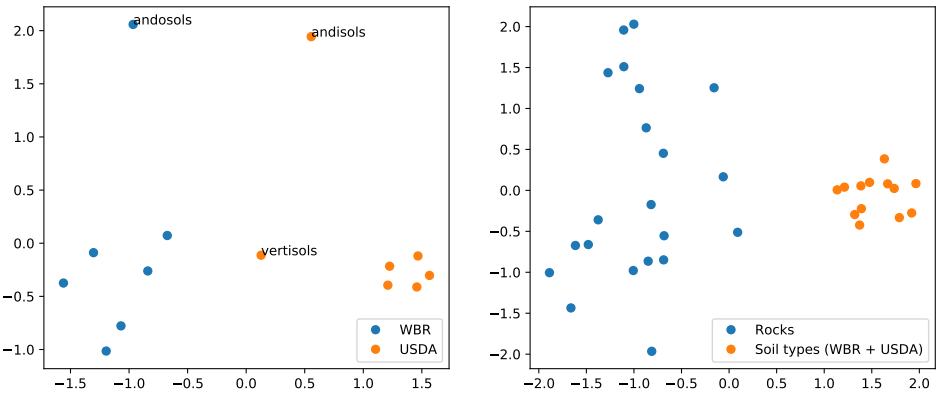


Fig. 6.6: Two-dimensional PCA projection of selected categorisations. Clusters representing soil types from different soil classification systems (left panel) and a different aggregation level where the same soil types are grouped as a single cluster when compared with rocks (right panel).

between text embeddings is less common. Bowman *et al.*, (2015) analysed the latent vector space of sentences and found that their model was able to generate coherent and diverse sentences when sampling between two embeddings. Duong *et al.*, (2016) interpolated between embedding from two vector spaces trained on different languages corpora to create a single cross-lingual vector space. The vector space from our model also presents similar characteristics.

In order to generate the interpolated embeddings, we obtained linear combinations of two words embeddings by using the formula

$$v_{int} = \alpha * v_a + (1 - \alpha) * v_b \tag{6.4}$$

where v_{int} is the interpolated embedding, v_a and v_b are the embeddings of the two selected words. By varying the value of α in the range $[0, 1]$, we generated a gradient of embeddings. For each intermediate embedding obtained by interpolation, we calculated the cosine similarity (Eq. 6.3) against all the words in the corpus and selected the closest one.

The results showed coherent concepts along the gradients (Fig. 6.7). The interpolation between "clay" and "boulder", with fine and coarse size, respectively, yields a gradient of sizes, with "clay" < "silt" < "sand" < "gravel" < "cobble" < "boulder". Another interpolation example, along another type of relationship, is shown in

the right panel of Fig. 6.7. The interpolation between the rocks "slate" and "migmatite" yields a gradient of rocks with different grades of metamorphism, with "slate" < "phyllite" < "schist" < "gneiss" < "migmatite".

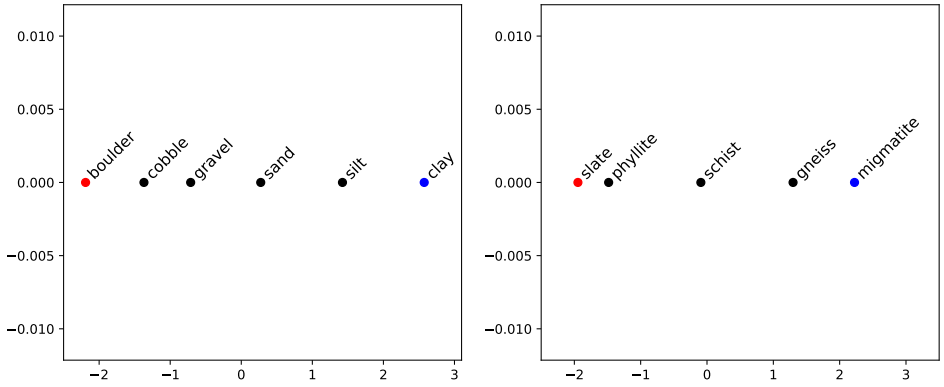


Fig. 6.7: Interpolated embedding in a two-dimensional PCA projection showing a size gradient (left panel) with "clay" < "silt" < "sand" < "gravel" < "cobble" < "boulder"; and gradient of metamorphism grade (right panel) with "slate" < "phyllite" < "schist" < "gneiss" < "migmatite". Red and blue dots represent selected words ("clay" and "boulder", and "slate" and "migmatite") and black dots represent the closest word (cosine similarity) to the interpolated embeddings.

6.6.6 Illustrative example

As a final, external evaluation of the embedding, we estimated average embeddings for each Great Group (Soil Taxonomy) of soils from 10,000 soil profiles descriptions. The convex-hulls at soil order level (Fig. 6.8) show the same pattern reported by Hughes *et al.*, (2017). Thanks to the unique characteristic of Histosols and the high diversity of this taxonomic group, they are easily differentiated in the 2-dimensional projection, showing the highest variability. The rest of the soil orders are heavily overlapped since their differences are hard to simplify into a 2-dimensional space. That overlap does not imply that the orders are not separable in a higher dimensional space. Here we are plotting the first 2 principal components (PCs), which only account for 28.8% of the total variance. This is probably the same reason for the overlap in the study by Hughes *et al.*, (2017) since they account for a 95% of the total variance only after 36 PCs (i.e. their plot, also using the first 2 PCs, probably explain a low proportion of the total variance, similar to our example).

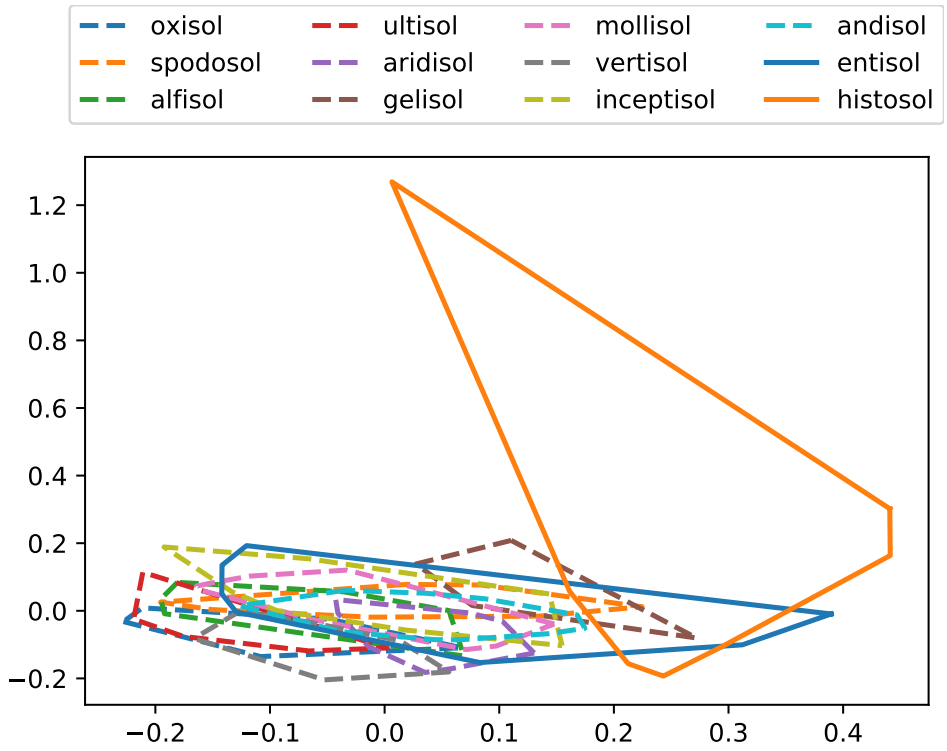


Fig. 6.8: Convex-hulls of great group embeddings at the order level (Soil Taxonomy). Great group embeddings were obtained after averaging the embeddings of all the words in the descriptions of the profiles belonging to each great group. The convex-hulls were estimated from the 2 first principal components of the great group embeddings.

This example shows how, by using descriptions encoded as word embeddings, we were able to use the same methods used by Hughes *et al.*, (2017). In this case, if no soil variables (laboratory data) were available, word embeddings could be used instead. Ideally, we would expect to use word embeddings to complement numerical data to utilise valuable information included in the descriptive data. This is also possible with other approaches. Hughes *et al.*, (2017) manually generated binary embeddings for the presence of ice and water. Another alternative to create embeddings is fuzzy logic. For example, McBratney and Odeh (1997) fuzzify categorical information from soil profiles such as depth, generating an encoding that represents the probability to belong to different depth classes (e.g a "fairly deep" soil could lay between the "shallow" and "deep" classes, with a membership of 0.5 to each class). The advantage of using word embeddings is that they are high dimensional vectors that encode much more

information applicable to many tasks, which would be difficult to replicate by manual encoding.

6.6.7 What do these embeddings actually represent?

It is worth discussing if word embeddings tell us anything about nature or if they really just tell us about the humanly constructed way that science is done and reported. A language model extracts information from the corpora to generate a representation in a high dimensional space. This continuous vector space shows interesting features that relate words to each other, which were tested in multiple tasks designed to evaluate the syntactic regularities encoded in the embeddings. Considering the position that science is a model of nature (Gilbert, 1991) and assuming that the way we do and report science is a good representation of it, if the language model is a good representation of the corpora of publications, perhaps the derived syllogism — the language model is a good representation of nature — can be considered as true. Of course, the representation of a representation carries many impressions, but it is worth exploring its validity.

As shown by the linear combinations of embeddings (Fig. 6.7), some aspects related to "size" are captured by the embeddings and, even if size categories are a human construct, they describe a measurable natural property. A more complex case is the illustrative example, where the embeddings capture some aspect of nature which are also captured by the numerical representation of its properties (in this case soil properties such as clay content, pH, etc). Given the results of the intrinsic evaluation of this work and others referenced throughout this article, it is probably impossible to generate the "perfect embeddings". Even if we were able to process all the written information available, and ignoring the limitations of any language model, the embeddings would be still limited by our capacity to understand non-linear relationships (Doherty and Balzer, 1988) and, in consequence, to understand nature.

Whether word embeddings can give new insights about geosciences is still to be tested. Studies in other fields have shown some potentially new information. For instance, Kartchner *et al.*, (2017) generated embeddings from medical diagnosis data and, after performing a clustering, they found clear links between some diagnoses related to advanced chronic kidney disease. Some of the relations are already known and accepted by the medical community while others are new and are just starting to

be studied and reported.

6.6.8 Future work

In the future, we expect to evaluate the effect of using our embeddings in more downstream applications (extrinsic evaluation). It is expected that domain-specific embeddings will necessarily improve the results of downstream tasks but this is not always the case. Schnabel *et al.*, (2015) suggested that extrinsic evaluation should not be used as a proxy for a general notion of embedding quality, since different tasks favour different embeddings, but they are useful in characterising the relative strengths of different models. We also expect to expand the test suite with more diverse and complex tests, opening the process to the scientific community. Another interesting opportunity is the inclusion of word embeddings in numerical classification systems (Bidwell and Hole, 1964; Crommelin and De Gruijter, 1973; Sneath and Sokal, 1973; Webster, 1977; Hughes *et al.*, 2014) which try to remove subjectivity by classifying an entity (soil, rock, etc.) based on numerical attributes that describe its composition.

6.7 Conclusions

In this work we introduced the use of domain-specific word embeddings for geosciences (GeoVec), and specifically soil science, as a way to a) reduce inconsistencies of descriptive data, and b) open the alternative to include such data into numerical data analysis. Comparing the result with general domain embeddings, trained on corpus such as Wikipedia, the domain-specific embedding performed better in common natural language processing tasks such as analogies, terms relatedness and categorisation, improving the overall accuracy by 107.9%.

We also presented a test suite, specifically designed for geosciences, to evaluate embedding intrinsic performance. This evaluation is necessary to test if syntactic or semantic relationships between words are captured by the embeddings. The test suite comprises tests for three tasks usually described in the literature (analogy, relatedness and categorisation) with different levels of complexity. Since creating a set of gold standard tests is not a trivial task, we consider this test suite a first approach. In the future, we expect to expand the test suite with more diverse and complex tests and to

open the process to the scientific community to cover different subfields of geosciences.

We demonstrated that the high-dimensional space generated by the language model encodes different types of relationships, through examples of soil-related concepts. These relationships can be used in novel downstream applications usually reserved for numerical data. One of these potential applications is the inclusion of embeddings in numerical classification. We presented an example where we successfully emulated part of a taxonomic analysis of soil profiles which was originally applied to soil numerical data. By encoding soil descriptions as word embeddings we were able to use the same methods used in the original application and obtain similar results. Ideally, we would expect to use word embeddings when no numerical data is available or to complement numerical data to include valuable information included in the descriptive data.

6.8 References

Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., *et al.*, (2017). Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ* 14: 1–19.

Baroni, M., Bernardi, R., Do, N.-Q., and chieh Shan, C. (2012). Entailment above the word level in distributional semantics. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics: pp. 23–32.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1: pp. 238–247.

Baxter, W. and ichi Anjyo, K. (2006). Latent doodle space. In: *Computer Graphics Forum*. Vol. 25. 3. Wiley Online Library: pp. 477–485.

Bengio, Y. (2008). Neural net language models. *Scholarpedia* 3 (1): 3881.

Bidwell, O. and Hole, F. (1964). Numerical taxonomy and soil classification. *Soil Science* 97 (1): 58–62.

Bird, S. and Loper, E. (2004). NLTK: the natural language toolkit. In: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics: p. 31.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Caté, A., Perozzi, L., Gloaguen, E., and Blouin, M. (2017). Machine learning as a tool for geologists. *The Leading Edge* 36 (3): 215–219.

Crommelin, R. D. and De Gruijter, J. (1973). *Cluster analysis applied to mineralogical data from the coversand formation in the Netherlands*. Tech. rep. Stichting voor Bodemkartering Wageningen.

Davies, M. and Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*: 1047–1051.

Doherty, M. E. and Balzer, W. K. (1988). Cognitive feedback. In: *Advances in psychology*. Vol. 54. Elsevier: pp. 163–197.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*.

FAO (1988). FAO/UNESCO Soil Map of the World. Revised legend, with corrections and updates. World Soil Resources Report 60.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on information systems* 20 (1): 116–131.

Fonseca, F. T., Egenhofer, M. J., Agouris, P., and Câmara, G. (2002). Using ontologies for integrated geographic information systems. *Transactions in GIS* 6 (3): 231–257.

Gilbert, S. W. (1991). Model building and a definition of science. *Journal of Research in Science Teaching* 28 (1): 73–79.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In: *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*. Association for Computational Linguistics: pp. 40–48.

Heimerl, F. and Gleicher, M. (2018). Interactive analysis of word vector embeddings. In: *Computer Graphics Forum*. Vol. 37. 3. Wiley Online Library: pp. 253–265.

Hsu, W.-N., Zhang, Y., and Glass, J. (2017). Learning latent representations for speech generation and transformation. *arXiv preprint arXiv:1704.04222*.

Hughes, P., McBratney, A. B., Huang, J., Minasny, B., Micheli, E., and Hempel, J. (2017). Comparisons between USDA Soil Taxonomy and the Australian Soil Classification System I: Data harmonization, calculation of taxonomic distance and inter-taxa variation. *Geoderma* 307: 198–209.

Hughes, P. A., McBratney, A. B., Minasny, B., and Campbell, S. (2014). End members, end points and extragrades in numerical soil classification. *Geoderma* 226: 365–375.

Jain, A., Kulkarni, G., and Shah, V. (2018). Natural language processing. *International Journal of Computer Sciences and Engineering* 6 (1): 161–167.

Jiang, Z., Li, L., Huang, D., and Jin, L. (2015). Training word embeddings for deep learning in biomedical text mining tasks. In: *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE: pp. 625–628.

Kartchner, D., Christensen, T., Humpherys, J., and Wade, S. (2017). Code2vec: Embedding and clustering medical diagnosis data. In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE: pp. 386–390.

Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers* 7 (1): 3–10.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553): 436–444.

Maxwell, A. E., Warner, T. A., and Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing* 39 (9): 2784–2817.

McBratney, A., Mendonça Santos, M. L., and Minasny, B (2003). On digital soil mapping. *Geoderma* 117 (1): 3–52.

McBratney, A. B. and Odeh, I. O. (1997). Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma* 77 (2-4): 85–113.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. Vol. 26: pp. 3111–3119.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., tau Yih, W., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*: pp. 746–751.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM* 38 (11): 39–41.

Mosavi, A., Ozturk, P., and wing Chau, K. (2018). Flood prediction using machine learning models: Literature review. *Water* 10 (11): 1536.

Nooralahzadeh, F., Øvrelid, L., and Lønning, J. T. (2018). Evaluation of Domain-specific Word Embeddings using Knowledge Resources. In: *Proceedings*

of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018): pp. 1438–1445.

Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N., and Fei, S. (2016). Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution* 7 (11): 1262–1272.

Pakhomov, S. V., Finley, G., McEwan, R., Wang, Y., and Melton, G. B. (2016). Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics* 32 (23): 3635–3644.

Pande, H. (2017). Effective search space reduction for spell correction using character neural embeddings. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Vol. 2: pp. 170–174.

Peckham, S. D. (2014). The CSDMS standard names: Cross-domain naming conventions for describing process models, data sets and their associated variables: 67–74.

Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics* 40 (3): 288–299.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*: pp. 1532–1543.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

Roy, A., Park, Y., and Pan, S. (2017). Learning Domain-Specific Word Embeddings from Sparse Cybersecurity Texts. *arXiv preprint arXiv:1709.07470*.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM* 8 (10): 627–633.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*: pp. 298–307.

Sneath, P. H. and Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification.* P. 573.

Suits, D. B. (1957). Use of dummy variables in regression equations. *Journal of the American Statistical Association* 52 (280): 548–551.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In: *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics: pp. 384–394.

Upchurch, P., Gardner, J. R., Pleiss, G., Pless, R., Snavely, N., Bala, K., and Weinberger, K. Q. (2017). Deep Feature Interpolation for Image Content Changes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Vol. 1. 2: pp. 7064–7073.

USDA, N. (2010). Keys to soil taxonomy. *Soil Survey Staff, Washington*.

Venugopalan, S., Hendricks, L. A., Mooney, R., and Saenko, K. (2016). Improving LSTM-based video description with linguistic knowledge mined from text. *arXiv preprint arXiv:1604.01729*.

Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM: pp. 448–456.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., and Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics* 87: 12–20.

Webster, R. *et al.*, (1977). *Quantitative and numerical methods in soil classification and survey*. Clarendon Press.: p. 269.

Yeh, R., Liu, Z., Goldman, D. B., and Agarwala, A. (2016). Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961*.

Chapter 7

Soil Organic Carbon Space-time Assessment at the Global Scale

Summary

Soil organic carbon (SOC) is a key component of functional ecosystems and crucial for food, soil, water and energy security. Assessing SOC has become part of many global initiatives such as GlobalSoilMap and FAO's Global Soil Organic Carbon Map. All the derived products are developed within the digital soil mapping (DSM) framework where legacy or new soil profile data, usually from different years, are used to generate a single map. The DSM framework has proven to be effective for many applications, however, the resulting products are usually static, which is an important drawback, especially for dynamic soil properties such as SOC. In this work, we introduce a semi-mechanistic model to assess SOC at a global scale and through time, using the largest SOC samples database to date. Our model generates a SOC baseline which is then propagated through time by keeping track of annual landcover changes obtained from remote sensing products. All the SOC losses and additions follow landcover dynamics and depend on temperature and precipitation. We estimated a global carbon stock in the first 30 cm of around 793.16 Pg with annual losses due to landcover change of 1.93 Pg SOC yr^{-1} between 2001 and 2016. The biggest losses are concentrated in the tropic and sub-tropical regions, accounting for almost 50% of the total loss (0.9 Pg yr^{-1}). The proposed modelling framework is flexible, allowing it to be updated as more

or better data becomes available.

Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. there are no other authors of the publication according to these criteria;

4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and

5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

Contributors	Statement of contribution
José Padarian *Signature:* José Padarian *Date:* April 2, 2020	Conceptualisation Data analysis Writing
Budiman Minasny	Writing
Alex McBratney	Writing

7.1 Introduction

Soil organic carbon (SOC) is an important component of the biosphere, controlling important physical, chemical and biological processes and their interactions. It is also a key component of soil quality and productivity (Tiessen *et al.*, 1994), a reason why it has been declared crucial for food, soil, water and energy security (McBratney *et al.*, 2014). SOC oxidation is also part of the path by which CO_2 fixed by land plants returns to the atmosphere (Schlesinger and Andrews, 2000). Due to its importance, assessing SOC stocks globally has become part of many initiatives. For instance, the *GlobalSoilMap* project (Arrouays *et al.*, 2014) aims to generate high resolution maps of key soil properties with global coverage using a bottom-up approach (each country is in charge of their national maps), where SOC is the subject of most studies published to date. Another example is the Global Soil Organic Carbon Map developed by FAO (FAO, 2017) by requesting its participant members national SOC maps. The International Soil Reference and Information Centre (ISRIC) also released global maps of multiple soil properties, including SOC (Hengl *et al.*, 2017). All the aforementioned products are developed within the digital soil mapping (DSM) framework (McBratney *et al.*, 2003) where legacy or new soil profile data, usually from different years, are used to generate a single map. The DSM framework has proven to be effective for many applications, however, the resulting products are usually static, which is an important drawback, especially for dynamic soil properties such as SOC. This issue has been recognised by many DSM studies (e;.g. Mulder *et al.*, (2016) and Padarian *et al.*, (2017)) and justified by the limited amount of soil data usually available.

The recent IPCC report (August 2019) mentioned that the expansion of agriculture and forestry area to support food availability for a growing population, has contributed to increasing GHG emissions, loss of natural ecosystems and declining biodiversity. The effect of local landcover change on SOC stocks has been studied extensively, including changes due to agriculture (Russel *et al.*, 1929; Mann, 1986; Davidson and Ackerman, 1993), agricultural land abandonment (Ihori *et al.*, 1995; Zou and Bashkin, 1998; Silver *et al.*, 2000), and change from native forest to plantations (Kasel and Bennett, 2007; Eclesia *et al.*, 2012). At the global scale, considering the usually limited amount of soil information, there is still no clear indication of how much soil carbon has been lost due to the expansion of agriculture and forestry area. Usually, studies rely on

coupled climate–carbon cycle general circulation models (Cox *et al.*, 2002) or numerical simulation with mechanistic models (Jones *et al.*, 2005). Also, it is possible to find studies using small soil datasets to extrapolate to the globe (Crowther *et al.*, 2016) that might lead to misleading results (Van Gestel *et al.*, 2018).

At the global scale, landcover changes have an important role on soil carbon stocks (Guo and Gifford, 2002) which, in conjunction with the dramatic increase in extent of human intervention (Watson *et al.*, 2014), makes necessary the integration of this dynamic component into SOC assessment. Since obtaining accurate and updated landcover information at global extent is challenging, many studies rely on the use of remote sensing optical data (Congalton *et al.*, 2014).

Here we introduce a semi-mechanistic model to assess SOC at global scale and through time, using the largest compilation of soil information to date. This represents an update to our previous work (Stockmann *et al.*, 2015) where our model included landcover as a covariate but without considering the complete landcover history, hence lacking memory and ignoring SOC dynamics. This yields undesirable results such as abrupt changes in the SOC content, especially in situations when the landcover transitions to a higher SOC density class (e.g. agriculture to forest). Our new model generates a SOC baseline which is then propagated through time by keeping track of annual landcover changes derived from satellite imagery, where all the SOC losses and additions follow dynamics dependent on temperature and precipitation.

7.2 Materials and methods

7.2.1 The data

The data used in this work correspond to a compilation of multiple regional, national and global databases. Most of these databases have been compiled as an effort to rescue valuable soil information which is usually fragmented and stored in formats not readily usable (Arrouays *et al.*, 2017). The complete dataset comprises 110,772 profiles with SOC information, most of them covering up to one meter depth, from which only 77,424 were used in our analysis since they report a sampling date. More details about the data can be found in our previous work (Stockmann *et al.*, 2015). The resulting 77,424 profiles (Fig. 7.1) were standarised to fixed depth intervals (0-5, 5-15,

15-30, 30-60, and 60-100 cm) using an equal-area spline algorithm (Bishop *et al.*, 1999), following the specifications of the *GlobalSoilMap* project. As explained in the following sections, we use satellite imagery to extract landcover information at each sampling location. Since the landcover product used in this study is available from the year 2001 onward, only 39,629 profiles had landcover information. From the observations corresponding to years before 2001, we only considered the ones that did not presented landcover changes between 2001 and 2017 (27,231 profiles) and we assumed the current landcover as a proxy.

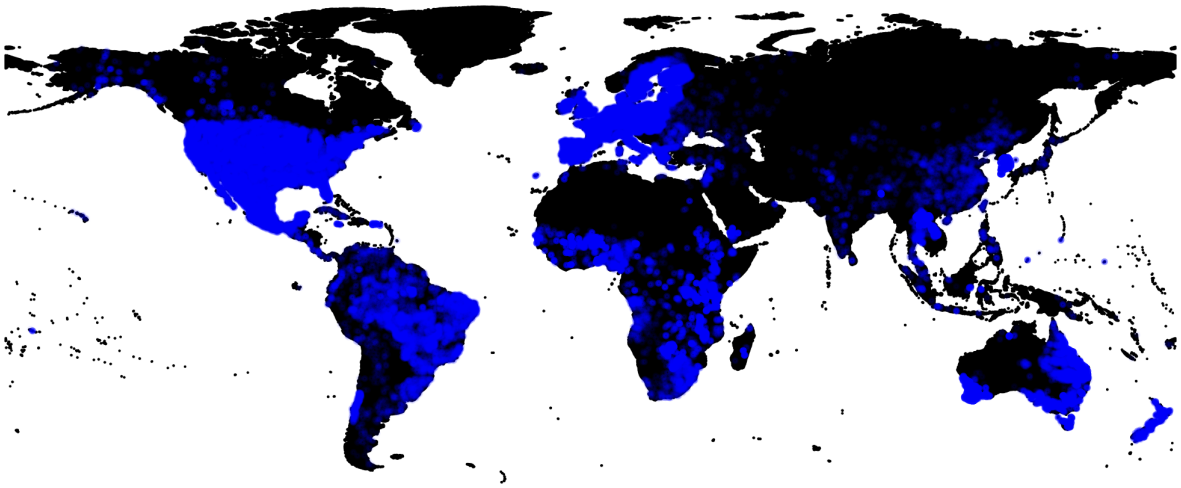


Fig. 7.1: Location of the 77,424 observations used in this work. Points are semi-transparent, hence intense blue areas have overlapping symbols.

To calculate the SOC stocks, and since the bulk density (BD) information in our dataset is limited, we used the pedotransfer function by Adams (1973) to predict BD based on the SOC content of the resulting maps.

$$BD = \frac{100}{\left(\frac{1.72xoc}{BD_{\text{organic}}}\right) + \left(\frac{100-1.72xoc}{BD_{\text{mineral}}}\right)} \tag{7.1}$$

where BD_{organic} and BD_{mineral} are the densities of the organic and mineral fractions, 0.223 and 1.32 respectively, and 1.72 is the factor used to convert from OC to organic matter content.

In this work we report different estimations (e.g. total stocks, losses) for different regions of the globe. One of them is the Global Ecoregions map, sourced from The

Nature Conservancy, USDA Forest Service and U.S. Geological Survey (Fig. 7.2).

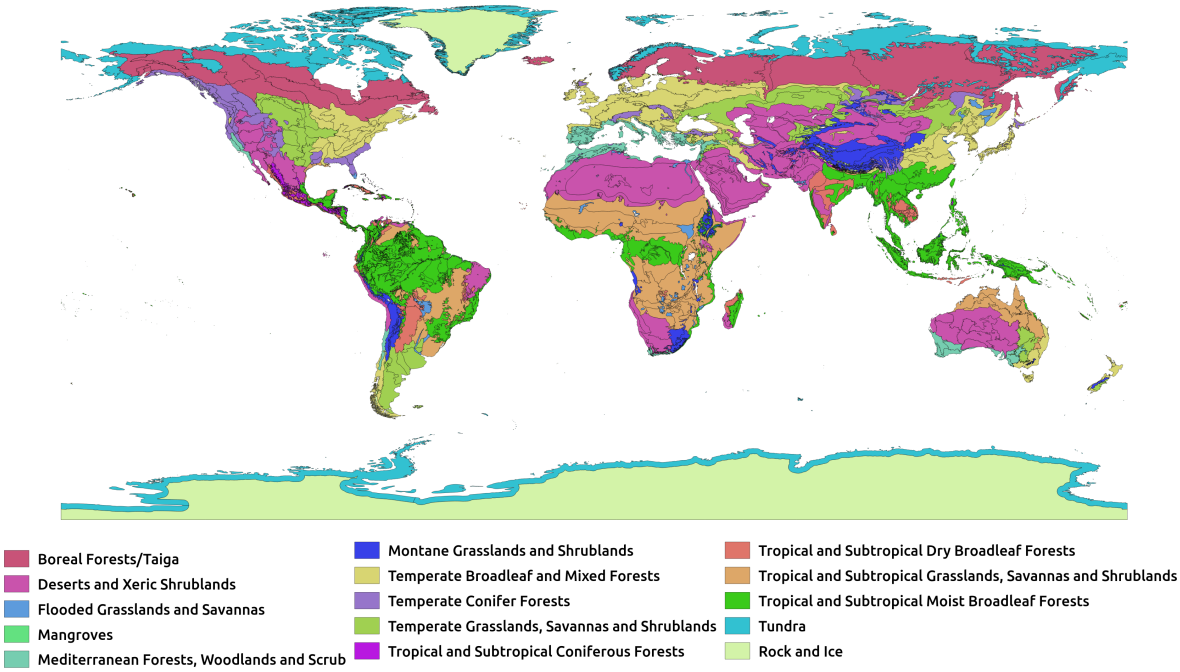


Fig. 7.2: Map showing Global Ecoregions used in this work (metadata can be found at http://maps.tcn.org/files/metadata/TerrEcos.xml).

7.2.2 Spatio-temporal model

We performed a multiple-stage modelling, mixing statistical and mechanistic inference. In brief, on the first stage we generated a "baseline" model for the year 2001, with the form

$$SOC_{xy} = f(\text{elevation}_{xy}, \text{slope}_{xy}, \text{MAT}_{xy}, \text{TAP}_{xy}, \text{landcover}_{xy}) \tag{7.2}$$

where MAT is the mean annual temperature and TAP the total annual precipitation, and none of the variables varies in time. To account for the time component, in the second stage we performed a "landcover tracking" routine starting from our baseline, to establish where and when landcover changes occur, the nature of the changes, and how long the new landcovers persist, yielding a 1-year temporal resolution series of maps. The details about both modelling stages are described in the

following sections. Both stages were carried out using a platform called *Google Earth Engine*, due to two main advantages, namely the availability of raster datasets used as covariates in our model and the capacity of performing at scale, parallel computations using Google's infrastructure (Padarian *et al.*, 2015).

Stage 1: SOC baseline generation

We based the first stage of our space-time modelling on the *scorpan* regression kriging approach (McBratney *et al.*, 2003), where a soil attribute (i.e.: SOC) is a function of a series of soil forming factors which are represented by environmental covariates (Eq. 7.2). In a digital soil mapping framework, these covariates are usually in the form of raster images. In this study we used the following set of covariates: a) 30 m digital elevation model (Danielson and Gesch, 2011) and slope (D-4), b) 1 km resolution long-term mean annual temperature (MAT) and total annual precipitation (TAP) (Hijmans *et al.*, 2005), and c) 500 m resolution land cover (MODIS product MCD12Q1.006). All the covariates were resampled to 500 m resolution. To link the environmental covariates with the SOC content we used the classification and regression tree (CART) algorithm (Breiman *et al.*, 1984).

Since the spatial coverage of the soil information for each year is limited and many landcover classes are under-represented, we used all the soil observations after extracting their landcover class from their corresponding years. This is equivalent to a time-for-space substitution, similar to the space-for-time approach commonly used in ecology as an alternative to long-term studies (Pickett, 1989).

Stage 2: Landcover tracking

We used the MODIS Land Cover Type product (MCD12Q1.006) generated by the Land Processes Distributed Active Archive Center, U.S. Department of the Interior and U.S. Geological Survey (DOI: 10.5067/MODIS/MCD12Q1.006). This product provides yearly global land cover information at 500 m resolution, from the year 2001 until 2017 (updated yearly), using five global land cover classification systems. In this study we used the International Geosphere Biosphere Programme (IGBP) classification system (Loveland and Belward, 1997; Belward, 1999).

Because some of the classes were insufficiently represented in our dataset, we merged

the 13 original classes into nine aggregated classes with comparable SOC densities using K-means clustering. The original classes and resulting new classes are shown in Fig. 7.3 and Table 7.1.

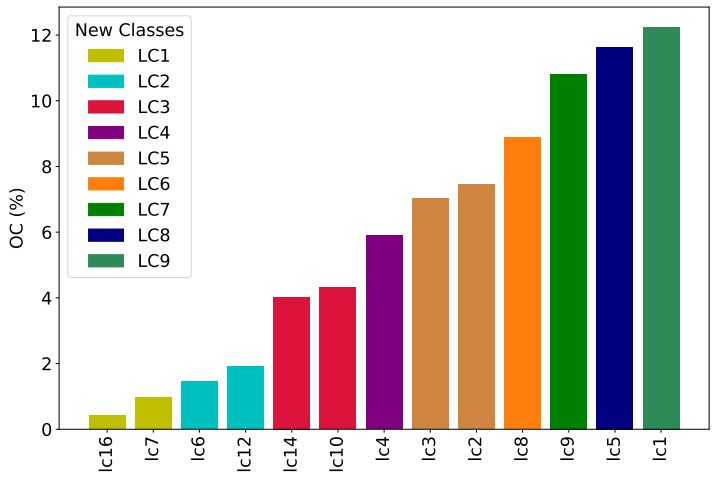


Fig. 7.3: Mean organic carbon content (%) of the new classes generated by a K-means clustering of original MODIS land cover classes. Labels in x axis correspond to the original land cover class (see Table 7.1).

We analysed each annual raster image since 2001 and kept track of all the landcover changes, per pixel, to establish if, in any of the consecutive years, a positive or negative SOC concentration change would occur. The changes in SOC follow the dynamics described in Section 7.2.3.

7.2.3 Soil organic carbon dynamics

Magnitude of change: $\widetilde{m}$

When a landcover change occurs, it triggers a series of events leading to changes in the properties of the affected system, including SOC content. These changes usually happen over a period of time until the system reaches a new equilibrium. In this work we rationalise that these equilibrium states are related to the mean SOC content of each landcover. We also considered that SOC dynamics due to landcover change vary depending on climate, and to account for this difference we used the Köppen-Geiger Climate Groups (Peel *et al.*, 2007) to aggregate our observations.

Table 7.1: Land cover simplification

Class name	MCD12Q1 class	New class
Evergreen Needleleaf forest	1	9
Evergreen Broadleaf forest	2	5
Deciduous Needleleaf forest	3	5
Deciduous Broadleaf forest	4	4
Mixed forest	5	8
Closed shrublands	6	2
Open shrublands	7	1
Woody savannas	8	6
Savannas	9	7
Grasslands	10	3
Croplands	12	2
Cropland/Natural vegetation mosaic	14	3
Barren or sparsely vegetated	16	1

Instead of using the absolute SOC mean values of each landcover class and climate grouping (Table 7.2), we estimated the proportional difference between landcover classes, as $\frac{new\ landcover}{previuos\ landcover}$. For example, it is expected that a change in landuse from landcover class 9 to landcover class 2 within a continental climatic zone, would result in a SOC reduction of 79% (1 - 2.54/12.26).

Rate of change: $\widetilde{r}$

The transition between the initial and the new system equilibrium state happens over a period of time and here, together with the magnitude of the change, the rate of that change is dependent on temperature and water availability.

The temperature dependence (v_t) is characterised by

$$v_t = e^{-E/k(T+273.15)}, \tag{7.3}$$

which is also known as the Arrhenius function (Arrhenius, 1915), where E corresponds to an "activation energy", T is the MAT in degrees Celsius, and k is

Table 7.2: Mean SOC content (in percentage) by landcover and Köppen-Geiger climate groups.

	Tropical	Dry	Temperate	Continental	Polar/Alpine
lc1	1.45	0.87	1.96	7.14	13.21
lc2	1.46	1.15	1.85	2.54	2.68
lc3	1.65	1.17	4.00	3.80	8.33
lc4	4.82	4.75	5.30	6.74	3.11
lc5	4.68	4.28	4.09	3.61	6.84
lc6	4.15	4.65	4.04	10.98	8.52
lc7	2.61	2.06	9.84	4.57	4.73
lc8	2.43	12.89	12.07	9.62	6.26
lc9	–	3.56	10.61	12.06	8.91

the Boltzmann's constant ($8.65 \times 10^{-5} eVK^{-1}$). By modifying the activation energy, this equation can be used to: *a*) describe SOC gains based on the temperature dependence of Rubisco carboxylation ($E \approx 0.32eV$), or *b*) describe SOC losses based on the temperature dependence of processes governed by respiration ($E \approx 0.65eV$). Despite the fact that Arrhenius function is generally used to explain the dependency on temperature at molecular level, it is also a central part of the "metabolic theory of ecology" where the Arrhenius formulation accounts for much of the variation in temperature dependence of biological processes at many levels (Brown and Sibly, 2012).

The dependence on precipitation (v_p) is described by a logistic function

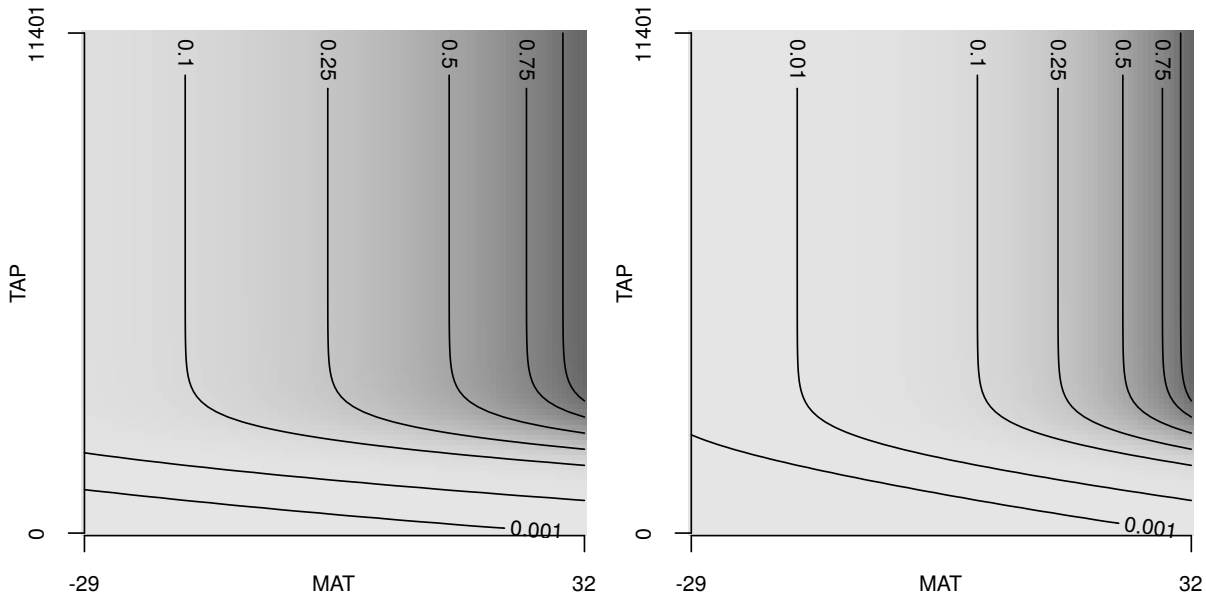
$$v_p = \frac{1}{1 + e^{0.003(tap - 11401/5)}} \tag{7.4}$$

where tap is the total annual precipitation in mm. In out model, we assume that at low precipitation levels SOC flows are negligible (Ewing *et al.*, 2008), and after certain amount of water has been stored in the soil, its influence can be dismissed if we assume free-draining conditions. Similar logic has been used by Van Veen and Paul (2011) who used soil moisture deficit as a reduction factor to simulate SOC dynamics in grassland soils.

The combination of both Eq. 7.3 and Eq. 7.4 yields to a normalised rate-modifying factor $\widetilde{r}$,

$$\widetilde{r} = \frac{r}{r_{max}},$$
$$r = v_t x v_p \tag{7.5}$$

where r_{max} represents the r value for the maximum values of temperature (32 °C) and precipitation (11,401 mm) on the rasters mentioned on Section 7.2.2. The values of r_{max}, for gain and losses are 5.43×10^{-6} and 2.02×10^{-11}, respectively. The value of $\widetilde{r}$ over the range of temperature and precipitation, for gain and losses, are depicted in Fig. 7.4).



(a) SOC gains (b) SOC losses

Fig. 7.4: Normalised rate-modifying factor ($\widetilde{r}$) surfaces for the range of temperature and precipitation used in this study.

Shape of change: $\widetilde{s}$

To account for SOC accumulation, we generated a sequence of weights $\widetilde{s}_{gain}$ using a logistic function, which length and shape depend on the value of the normalised

rate-modifying factor $\widetilde{r}$ (Eq. 7.5).

$$\widetilde{s}_{gain} = \frac{s_{gain}}{\sum s_{gain}}$$

$$s_{gain} = \left(\frac{1}{1 + e^{\left(-c\, l^{-b} \left(i - \frac{l}{2}\right)\right)}} \right)_{i=1}^{l} \tag{7.6}$$

$$l = a + floor(\widetilde{r}^{-b}),$$

where l is the number of years to reach a new equilibrium and a, b and c are shape parameters. A logistic function is a common way to describe growth. Richards' equation (Richards, 1959), and Gompertz's function (Gompertz, 1825) are some examples. This type of function has also been applied to predict growth in forest ecosystems (Botkin, 1993; Pacala *et al.*, 1993), crops (Yin *et al.*, 2003), and to explain SOC accumulation (Knops and Tilman, 2000; Huang *et al.*, 2012).

To simulate the SOC loss mechanism, we used an exponential-decay function to generate a sequence of weights $\widetilde{s}_{loss}$, which also depends on Eq. 7.5.

$$\widetilde{s}_{loss} = \frac{s_{loss}}{\sum s_{loss}}$$

$$s_{loss} = \left((1 - f)^{i} \right)_{i=1}^{l} \tag{7.7}$$

$$l = d + floor(\widetilde{r}^{-e}),$$

where d, e and f are shape parameters. This type of curve has been widely used to describe SOC decomposition (Covington, 1981; Poeplau *et al.*, 2011). Both Eq. 7.6 and Eq. 7.7 have three shape parameters that dictate the corresponding SOC gain o loss, which were optimised in two separate Markov Chain Monte Carlo (Gelfand *et al.*, 1990) routines. The prior distributions were generated using information (i.e. temperature, precipitation, years to equilibrium) extracted from the literature (Table 7.A.1). Both

optimisations were run for 200,000 steps, using a "burn-in" period of 120,000 steps. The optimised parameters corresponded to: $a = 80$, $b = 0.7$, $c = 2.5$, $d = 5$, $e = 0.4$, and $f = 0.2$.

Total change

By combining magnitude (Section 7.2.3), rate (Section 7.2.3) and shape (Section 7.2.3), it is possible to assess the differences in SOC for any given location after a landcover change. For instance, if we consider a location belongs to a continental climatic group, that experienced a change from landcover class 9 to a landcover class 2 ($\widetilde{m} = 0.21$), with $MAT = 24^{o}C$ and $TAP = 1900mm$ ($\widetilde{r} = 0.125$), using Eq. 7.7 we can obtain a loss sequence $\widetilde{s}$ = (0.253, 0.202, 0.162, 0.13, 0.104, 0.083, 0.066). By computing $\widetilde{s} \times \widetilde{m}$ we obtain a sequence of loss (proportion of initial SOC concentration) for each year until the new equilibrium is reached (7 years later).

7.2.4 Model evaluation

In order to validate the model, we performed a bootstrapping routine (Efron and Tibshirani, 1993) where the dataset was sampled with replacement before fitting the model. Due to computation limitations, we only fitted 10 models for each depth interval. At each of the iterations, the model was evaluated against observations not included on the bootstrap sample by calculating the coefficient of determination (R^2), the root of the mean squared error (RMSE) and the root of the median squared error (RMdSE). We decided to include RMdSE since the SOC distribution is heavily right skewed and a measure based on the median squared error is a more robust alternative (Kempen *et al.*, 2012) to evaluate the overall performance of our model.

7.3 Results and discussion

7.3.1 Baseline evaluation

The baseline model for the year 2001 shows an expected performance, with a mean R^2 of 0.52 and a RMSE of 9.44 SOC % . Similar values have been reported by other DSM studies at global scale (e.g. Hengl *et al.*, (2017)). Considering the wide range of

SOC values, RMSE could be a misleading measure since the largest errors are due to the underestimation of high carbon values. In this case, the RMdSE for the baseline model is 0.95%.

The prediction map (mean of the 10 bootstrapping iterations; Fig 7.5) shows the expected SOC distribution patterns with low SOC contents in areas such as the Sahara, Arabian, Gobi, Australian and Atacama deserts, and high SOC contents towards the poles. In terms of the uncertainty (Fig 7.6), the spatial pattern has the expected behaviour with higher uncertainties in areas either with large SOC contents (Poggio *et al.*, 2013) or poorly represented in our dataset (e.g. The Andes; (Padarian *et al.*, 2017)).

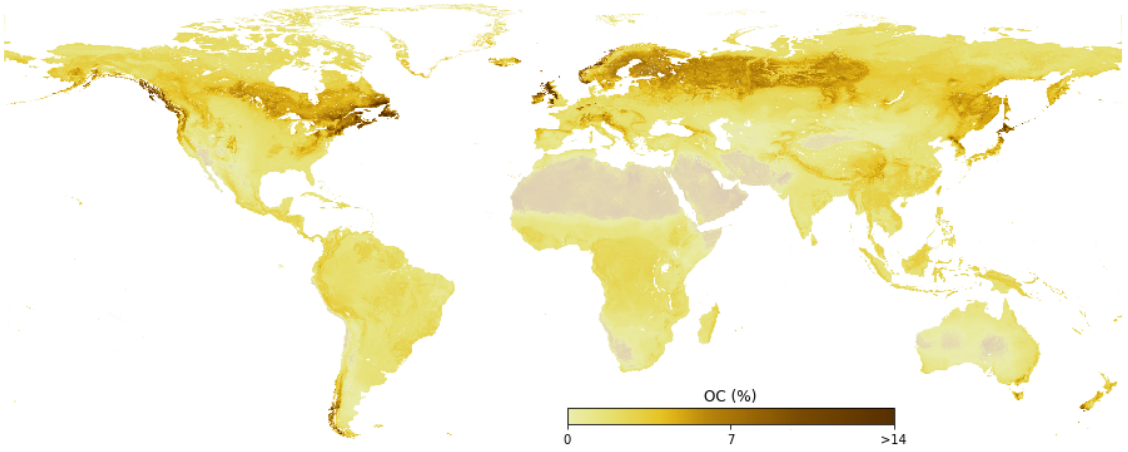


Fig. 7.5: Map of predicted SOC (%). The map corresponds to the mean value of the 10 bootstrapping iterations, for the 0-5 cm depth interval.

7.3.2 Temporal evaluation

For the temporal validation, we used the samples not selected during each iteration of the bootstrapping routine. The performance in time was acceptable, with a mean RMSE of 8.72% and mean RMdSE of 0.99%. These values are similar to those obtained during the baseline prediction, indicating that the landcover tracking routine of our model did not introduce new error. However, it is important to consider that the baseline and many of the components of the time tracking routine have uncertainty associated with them, which should be propagated in space and time. Due to the

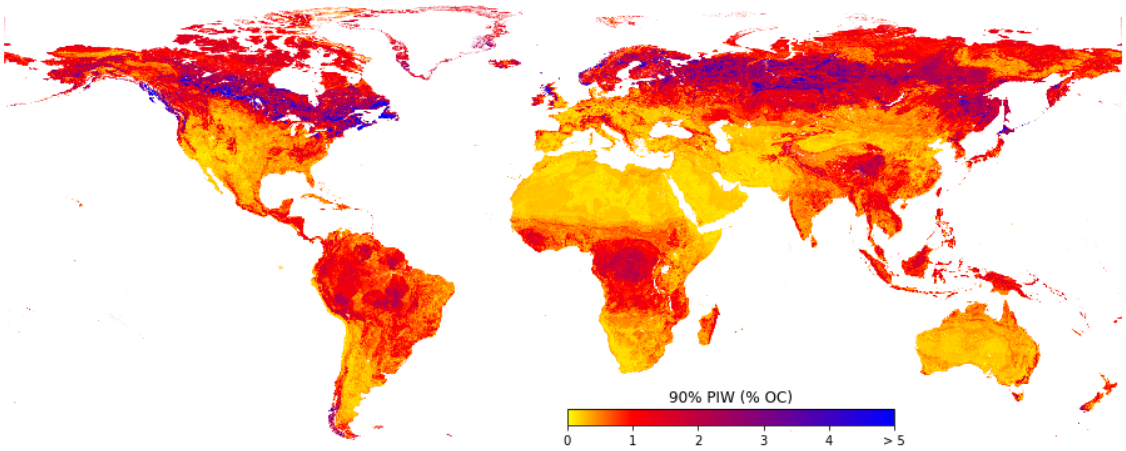


Fig. 7.6: Uncertainty map. The map corresponds to the 90% prediction interval width of the 10 bootstrapping iterations, for the 0-5 cm depth interval.

computational burden of this task, we will perform a complete uncertainty assessment in a future project.

Thanks to the landcover tracking, the model is able to consider the landcover history of an area and behaves as expected. For instance, Fig.7.7 depicts the transition from native forest to an oil palm tree plantation in Borneo. Initially, the area was considered as native forest, and at some point during 2008-2009 the palm tree plantation was established. For the following 3-4 years, while the trees were small, MODIS classified that area as cropland and the model returned a lower SOC content. In 2013, when the trees were fully grown, the area was once again considered as native forest, even if the area was still a plantation. Thanks to the landcover tracking, the SOC values returned by the model did not increase to the original levels.

The global SOC content change is predominantly dominated by losses and are present all around the globe (Fig. 7.8). Most of the more evident changes are related to the contraction or fragmentation of forest, including the borders of the Congo basin and Amazon basin, and the limits of the natural forest formations in Australia and China.

More in detail, it is possible to observe a greater diversity of dynamics, not distinguishable in a global map. For instance, Fig. 7.9 shows an example in Rondonia State, Brazil, where losses as a result of the conversion from native forest to agricultural land are evident. It is also possible to observe that agricultural land established before

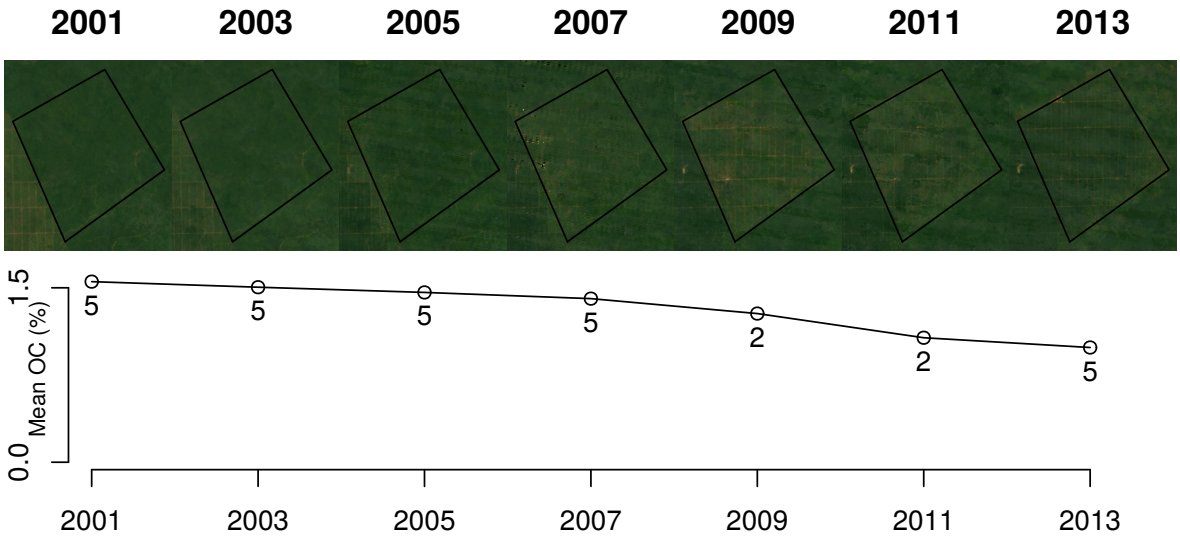


Fig. 7.7: Time series for ≈2,100 ha. in Borneo (117.0583° longitude, 1.15675° latitude) showing a transition from native forest to palm tree plantation. Upper panel: sequence of Landsat 7 imagery (annual composite to remove clouds). Lower panel: mean SOC (%) content, and the labels under the points correspond to the most common landcover class within the delimited area, according to Table 7.1.

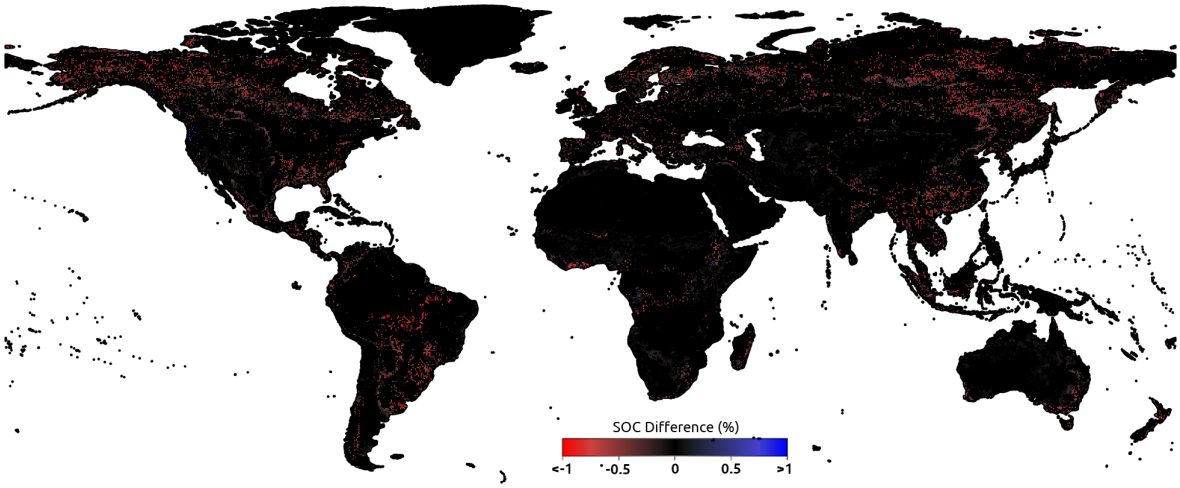


Fig. 7.8: Global topsoil SOC loss between 2001 and 2016. The colour is a gradient from -1% loss (red) to 1% gain, with no change as black. Values outside that range are shown in bright red (losses ≤ −1%) and bright blue (gains ≥ 1%).

our baseline year (2001) presents some modest gains as a consequence of the succession forest → bare soil → pasture which is very common in that area (de Moraes *et al.*,

1996).

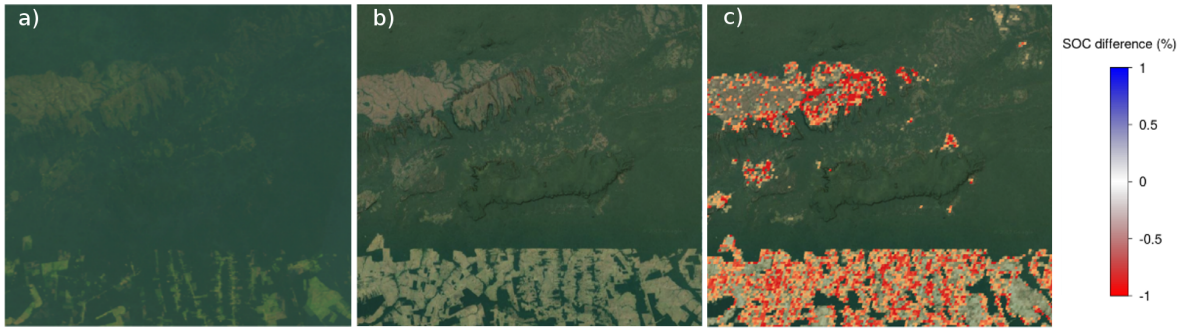


Fig. 7.9: Topsoil SOC loss in Rondonia State, Brazil. a) Satellite image for the year 2001; b) satellite image for year 2016; c) soil SOC for the year 2016 as compared to 2001.

7.3.3 Global stocks

According to our estimates, the topsoil (0-30 cm) SOC stock for the baseline year (2011) is around 793.16 Pg C which is close to the median value for the first meter (1,437 Pg C) of seven spatially-explicit studies reported by Scharlemann *et al.*, (2014) if we consider that the top 30 cm accounts for around 50% of the total stock up to 1 m (Jobbágy and Jackson, 2000). Our estimations are also slightly larger compared to values reported by Batjes (1996), with 684-724 Pg in the upper 30 cm. This is probably not an over-estimation since our database is the most detailed to date, including many samples from areas poorly represented in previous studies. For instance, Batjes (1996), used the WISE soil database that comprises very few profiles for North America, Oceania and the north temperate regions (Scharlemann *et al.*, 2014).

In terms of SOC changes in time (Table 7.3), our model indicates that there is a mean reduction of 1.93 Pg SOC yr^{-1} between 2001 and 2016. The biggest losses are concentrated in the tropical and sub-tropical regions, accounting for almost 50% of the total loss (0.9 Pg yr^{-1}).

Our estimates are larger than existing studies. For instance, Lal (2007) reports losses of around 1.6 Pg SOC yr^{-1} from the first 1-metre due to deforestation, land-use change and soil cultivation. This disagreement is partly due to our higher initial SOC estimates and also the finer resolution (500 m) of the landcover product we are using

Table 7.3: Mean annual soil organic carbon losses by ecoregion.

	Loss (Million tons C yr^{-1})
Tropical and Subtropical Moist Broadleaf Forests	-503.45
Tropical and Subtropical Dry Broadleaf Forests	-46.05
Tropical and Subtropical Coniferous Forests	-13.92
Temperate Broadleaf and Mixed Forests	-2.74
Temperate Conifer Forests	-53.66
Boreal Forests/Taiga	-40.38
Tropical and Subtropical Grasslands, Savannas a...	-335.10
Temperate Grasslands, Savannas and Shrublands	-148.58
Flooded Grasslands and Savannas	-97.98
Montane Grasslands and Shrublands	-15.27
Tundra	-62.59
Mediterranean Forests, Woodlands and Scrub	-228.21
Deserts and Xeric Shrublands	-287.33
Mangroves	-98.84

that allow us to capture more changes than the 1 km resolution used in other studies.

7.3.4 The effect of crop production

Agriculture has in important impact on SOC stocks, where management practices can make the difference between the soil becoming a sink or a source of carbon (Lal *et al.*, 2004). Since soil conservation practices are not widely implemented around the globe (Lal *et al.*, 2004), croplands generally behave as source of carbon into the atmosphere. If we only consider landcover change into cropland (defined by MODIS as lands covered with temporary crops followed by harvest and a bare soil period), with at least 5 years as such, the accumulated global loss since 2001 is around 1.41 Pg of SOC, with an average loss of 16.70 tons OC ha^{-1}. These values agree with the losses reported by Lal (2004), who also estimates that the conversion of natural to agricultural ecosystems can generate carbon losses as high as 20 to 80 tons C ha^{-1}.

7.3.5 The effect on productivity

SOC is crucial in agricultural environments since it affects processes such as the cation exchange between soil colloids and the soil solution (Parfitt *et al.*, 1995), provides energy for beneficial soil organisms (Ramesh *et al.*, 2019) and improves water holding capacity (Khaleel *et al.*, 1981). Agricultural productivity has been directly related with SOC contents and if reduced below some critical limits, soil condition declines affecting yields. For tropical soils, Aune and Lal (1997) reported that where SOC concentrations fall below a 1.1% critical limit, crop yields are reduced by 20%. Similarly, in temperate regions, a review by Loveland and Webb (2003) reported a critical limit of 2%.

According to our estimations, on average, each year in the tropical region around 11.07 million new hectares (110,692 km^2) fall below the 1.1% SOC critical limit proposed by Aune and Lal (1997). The surface affected increases by 247,680 ha yr^{-1}, consequence of the continuous anthropic pressure (Hansen *et al.*, 2013; Kim *et al.*, 2015). In temperate regions, on average, each year 5.36 million new hectares (53,622 km^2) fall below the 2% critical limit reported by Loveland and Webb (2003). There is a negative trend in the surface affected, with a 125,776 ha yr^{-1} decrease in the number of hectares that go under the 2% threshold.

Many of the new areas falling below critical limits were already under agricultural management in our baseline year. This highlights the need to change conventional management practices and, more generally, the food production chain. Continuing the trends of decreasing productivity of the land implies that new land has to be converted to compensate for the loss of production, and considering that a) the degradation of soils is usually fast, b) soil recovery is a slow process and c) productive land is a limited resource, we could face a serious food and environmental security problem.

7.3.6 Discounting landcover effect

The proposed model does not explicitly account for changes in temperature and precipitation over time or other sources of variation that might affect SOC stocks, such as leaching and erosion. Nevertheless, we wanted to explore the SOC trends discounting the effect of landcover change by selecting soil observations from locations where no landcover changes were detected by the MODIS product. We observed a general trend of decreasing SOC content in most areas and landcover combinations (Fig. 7.10) but

it is important to consider that the number of samples in some of them is limited and that the sampling coverage and location usually change over time, hence adding bias. For croplands (landcover class 2), management practices can have a confounding effect, especially considering the negative effect on soil carbon under conventional agriculture that might affect its condition even under a constant landcover class.

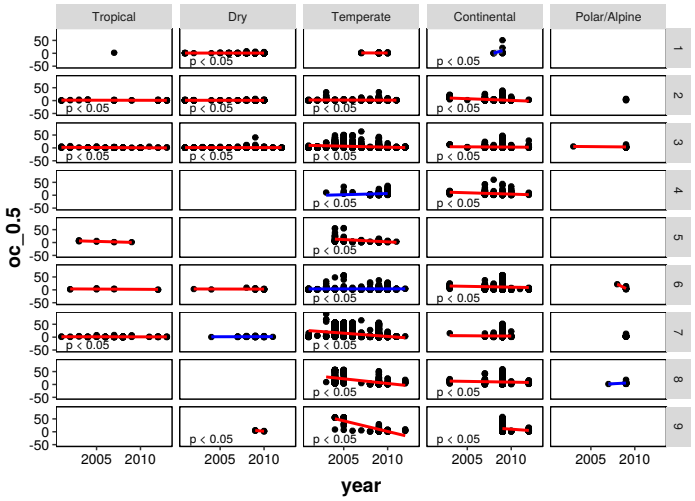


Fig. 7.10: Temporal trend in SOC changes between 2001 and 2017. To account for the effect of climate change, only observations in areas without landcover changes were included. Red and blue trend lines represent negative and positive slopes. Panels are annotated with "$p < 0.05$" when the slope is significantly different from zero.

7.3.7 Limitations

Despite the contribution towards a better assessment of SOC by incorporating a time component of this work, there are still many possible improvements including:

- Using soil texture information since it is a factor that influences carbon dynamics (Stockmann *et al.*, 2013), but, at the moment, there are no reliable global estimates.

- Including other auxiliary soil information relevant when estimating carbon stocks such as coarse fragments, soil thickness and mineral fraction density. Similar to soil texture information, there are no reliable global estimates.

- In our model, the SOC content is mainly governed by photosynthesis activity (accumulation) and microbial respiration (loss). We excluded processes that have shown to have a significant effect on the carbon balance such as leaching (Kindler *et al.*, 2011) and erosion (Lal, 2003; Berhe *et al.*, 2008).
- Many areas have a limited amount of data. This project has an ongoing component to try to add collaborators and new data and we are focusing our efforts to increase the representativeness of the dataset, both spatially and temporally.
- Another obvious limitation is that our model relies on satellite products which are only recently available. Specifically, the MODIS products are available since 2001, which limits the scope of our study considerably. A feasible solution that we plan to implement is to use Landsat imagery to downscale the MODIS product (Zhang and Roy, 2017). This will allow us to potentially reach a resolution of 30 m and, more importantly, set a new baseline in the year 1984. We designed the modelling framework with these limitations and potential improvements in mind. It is flexible enough to be easily updated as new or better data (soil, covariates, landcover) becomes available.

Conclusions

We successfully applied a semi-mechanistic model that includes a) a machine learning model to link environmental covariates with the largest soil database used to date, b) a mechanistic component that emulates carbon dynamics dependent on precipitation and temperature, and c) a landcover-tracking component that varies the outputs depending on the landcover history.

We estimated a global carbon stock in the first 30 cm of around 793.16 Pg with annual losses due to landcover change in the order of 1.93 Pg SOC yr^{-1} between 2001 and 2016. The biggest losses are concentrated in the tropical and sub-tropical regions, accounting for almost 50% of the total loss (0.9 Pg yr^{-1}).

The global carbon losses have a significant impact on productivity considering critical SOC values reported in other studies. Each year, in the tropical region around 11.07 million new hectares fall below the 1.1% SOC critical limit. The surface affected

increases by 247,680 ha yr^{-1}, as a consequence of the continuous anthropic pressure. In temperate regions, on average, each year 5.36 million new hectares fall below the 2% critical limit. There is a negative trend in the surface affected, with a 125,776 ha yr^{-1} decrease in the number of hectares that go under the 2% threshold.

Some important aspects related to soil carbon stocks such as stoniness and soil depth were not included in this model due to issues of data availability. Nevertheless, the proposed framework is flexible allowing it to be updated as more or better data becomes available.

7.4 References

Adams, W. (1973). The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils. *Journal of Soil Science* 24 (1): 10–17.

Arrhenius, S. *et al.*, (1915). Quantitative laws in biological chemistry.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., *et al.*, (2014). GlobalSoilMap: Toward a fine-resolution global grid of soil properties. In: *Advances in agronomy*. Vol. 125. Elsevier: pp. 93–134.

Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., *et al.*, (2017). Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ* 14: 1–19.

Aune, J. and Lal, R (1997). Agricultural productivity in the tropics and critical limits of properties of Oxisols, Ultisols, and Alfisols. *TROPICAL AGRICULTURE-LONDON THEN TRINIDAD-* 74: 96–103.

Batjes, N. H. (1996). Total carbon and nitrogen in the soils of the world. *European journal of soil science* 47 (2): 151–163.

Belward, A. S. (1999). The IGBP-DIS global 1-km land-cover data set DIS-Cover: A project overview. *Photogrammetric Engineering and Remote Sensing* 65: 1013–1020.

Berhe, A. A., Harden, J. W., Torn, M. S., and Harte, J. (2008). Linking soil organic matter dynamics and erosion-induced terrestrial carbon sequestration at different landform positions. *Journal of Geophysical Research: Biogeosciences* 113 (G4).

Bishop, T., McBratney, A., and Laslett, G. (1999). Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma* 91 (1): 27–45.

Botkin, D. B. (1993). *Forest dynamics: an ecological model*. Oxford University Press.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brown, J. H. and Sibly, R. M. (2012). The metabolic theory of ecology and its central equation. *Metabolic ecology: a scaling approach*: 21–33.

Congalton, R., Gu, J., Yadav, K., Thenkabail, P., and Ozdogan, M. (2014). Global land cover mapping: A review and uncertainty analysis. *Remote Sensing* 6 (12): 12070–12093.

Covington, W. W. (1981). Changes in forest floor organic matter and nutrient content following clear cutting in northern hardwoods. *Ecology*: 41–48.

Cox, P. M., Betts, R. A., Betts, A, Jones, C. D., Spall, S. A., and Totterdell, I. J. (2002). Modelling vegetation and the carbon cycle as interactive elements of the climate system. In: *International Geophysics*. Vol. 83. Elsevier: pp. 259–279.

Crowther, T. W., Todd-Brown, K. E., Rowe, C. W., Wieder, W. R., Carey, J. C., Machmuller, M. B., Snoek, B., Fang, S., Zhou, G., Allison, S. D., *et al.*, (2016). Quantifying global soil carbon losses in response to warming. *Nature* 540 (7631): 104.

Danielson, J. J. and Gesch, D. B. (2011). *Global multi-resolution terrain elevation data 2010 (GMTED2010)*. Tech. rep. US Geological Survey.

Davidson, E. A. and Ackerman, I. L. (1993). Changes in soil carbon inventories following cultivation of previously untilled soils. *Biogeochemistry* 20 (3): 161–193.

De Moraes, J. F., Volkoff, B., Cerri, C. C., and Bernoux, M. (1996). Soil properties under Amazon forest and changes due to pasture installation in Rondônia, Brazil. *Geoderma* 70 (1): 63–81.

Eclesia, R. P., Jobbagy, E. G., Jackson, R. B., Biganzoli, F., and Piñeiro, G. (2012). Shifts in soil organic carbon for plantation and pasture establishment in native forests and grasslands of South America. *Global Change Biology* 18 (10): 3237–3251.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Vol. 57. New York: CRC press: p. 436.

Ewing, S., Macalady, J., Warren-Rhodes, K, McKay, C., and Amundson, R (2008). Changes in the soil C cycle at the arid-hyperarid transition in the Atacama Desert. *Journal of Geophysical Research: Biogeosciences (2005–2012)* 113 (G2).

FAO (2017). Fifth Meeting of the Global Soil Partnership Plenary Assembly. . Available at: http://www.fao.org/3/a-bs973e.pdf (last access: 30 July 2019).

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* 85 (412): 972–985.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*: 513–583.

Guo, L. B. and Gifford, R. (2002). Soil carbon stocks and land use change: a meta analysis. *Global change biology* 8 (4): 345–360.

Hansen, M., Potapov, P., Moore, R, Hancher, M, Turubanova, S., Tyukavina, A, Thau, D, Stehman, S., Goetz, S., Loveland, T., *et al.*, (2013). High-resolution global maps of 21st-century forest cover change. *Science* 342 (6160): 850–853.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., *et al.*, (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PloS one* 12 (2): e0169748.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* 25 (15): 1965–1978.

Huang, L., Liu, J., Shao, Q., and Xu, X. (2012). Carbon sequestration by forestation across China: Past, present, and future. *Renewable and Sustainable Energy Reviews* 16 (2): 1291–1299.

Ihori, T., Burke, I. C., Lauenroth, W. K., and Coffin, D. P. (1995). Effects of cultivation and abandonment on soil organic matter in northeastern Colorado. *Soil Science Society of America Journal* 59 (4): 1112–1119.

Jobbágy, E. G. and Jackson, R. B. (2000). The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological applications* 10 (2): 423–436.

Jones, C., McConnell, C., Coleman, K., Cox, P., Falloon, P., Jenkinson, D., and Powlson, D. (2005). Global climate change and soil carbon stocks; predictions from two contrasting models for the turnover of organic carbon in soil. *Global Change Biology* 11 (1): 154–166.

Kasel, S. and Bennett, L. T. (2007). Land-use history, forest conversion, and soil organic carbon in pine plantations and native forests of south eastern Australia. *Geoderma* 137 (3-4): 401–413.

Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G., and de Vries, F. (2012). Efficiency comparison of conventional and digital soil mapping for updating soil maps. *Soil Science Society of America Journal* 76 (6): 2097–2115.

Khaleel, R, Reddy, K., and Overcash, M. (1981). Changes in soil physical properties due to organic waste applications: a review 1. *Journal of Environmental Quality* 10 (2): 133–141.

Kim, D.-H., Sexton, J. O., and Townshend, J. R. (2015). Accelerated deforestation in the humid tropics from the 1990s to the 2000s. *Geophysical Research Letters* 42 (9): 3495–3501.

Kindler, R., Siemens, J., Kaiser, K., Walmsley, D. C., Bernhofer, C., Buchmann, N., Cellier, P., Eugster, W., Gleixner, G., Grũnwald, T., *et al.*, (2011). Dissolved carbon leaching from soil is a crucial component of the net ecosystem carbon balance. *Global Change Biology* 17 (2): 1167–1185.

Knops, J. M. and Tilman, D. (2000). Dynamics of soil nitrogen and carbon accumulation for 61 years after agricultural abandonment. *Ecology* 81 (1): 88–98.

Lal, R. (2003). Soil erosion and the global carbon budget. *Environment international* 29 (4): 437–450.

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science* 304 (5677): 1623–1627.

Lal, R. (2007). Carbon sequestration. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1492): 815–830.

Lal, R., Griffin, M., Apt, J., Lave, L., and Morgan, M. G. (2004). *Managing soil carbon.*

Loveland, P and Webb, J (2003). Is there a critical level of organic matter in the agricultural soils of temperate regions: a review. *Soil and Tillage research* 70 (1): 1–18.

Loveland, T. R. and Belward, A. (1997). The international geosphere biosphere programme data and information system global land cover data set (DISCover). *Acta Astronautica* 41 (4-10): 681–689.

Mann, L. (1986). Changes in soil carbon storage after cultivation. *Soil Science* 142 (5): 279–288.

McBratney, A., Mendonça Santos, M. L., and Minasny, B (2003). On digital soil mapping. *Geoderma* 117 (1): 3–52.

McBratney, A., Field, D. J., and Koch, A. (2014). The dimensions of soil security. *Geoderma* 213: 203–213.

Mulder, V. L., Lacoste, M., de Forges, A. R., and Arrouays, D. (2016). GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth. *Science of the total environment* 573: 1352–1369.

Pacala, S. W., Canham, C. D., and Silander Jr, J. (1993). Forest models defined by field measurements: I. The design of a northeastern forest simulator. *Canadian Journal of Forest Research* 23 (10): 1980–1988.

Padarian, J, Minasny, B, and McBratney, A. B. (2015). Using Google's web-based platform for digital soil mapping. *Computers & Geosciences* 83: 80–88.

Padarian, J, Minasny, B, and McBratney, A. (2017). Chile and the Chilean soil grid: a contribution to GlobalSoilMap. *Geoderma Regional*: 17–28.

Parfitt, R., Giltrap, D., and Whitton, J. (1995). Contribution of organic matter and clay minerals to the cation exchange capacity of soils. *Communications in soil science and plant analysis* 26 (9-10): 1343–1355.

Peel, M. C., Finlayson, B. L., and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions* 4 (2): 439–473.

Pickett, S. T. (1989). Space-for-time substitution as an alternative to long-term studies. In: *Long-term studies in ecology*. Springer: pp. 110–135.

Poeplau, C., Don, A., Vesterdal, L., Leifeld, J., Van Wesemael, B., Schumacher, J., and Gensior, A. (2011). Temporal dynamics of soil organic carbon after land-use change in the temperate zone–carbon response functions as a model approach. *Global change biology* 17 (7): 2415–2427.

Poggio, L., Gimona, A., and Brewer, M. J. (2013). Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. *Geoderma* 209: 1–14.

Ramesh, T., Bolan, N. S., Kirkham, M. B., Wijesekara, H., Freeman II, O. W., Korres, N. E., Burgos, N. R., Travlos, I., Vurro, M., Salas-Perez, R., *et al.*, (2019). Soil organic carbon dynamics: Impact of land use changes and management practices: A review. In: *Advances in Agronomy*. Academic Press Inc.

Richards, F. (1959). A flexible growth function for empirical use. *Journal of experimental Botany* 10 (2): 290–301.

Russel, J., Harrison, C., and Wright, A. (1929). Organic matter problems under dry-farming conditions. *Agronomy journal* 21 (10): 994–1000.

Scharlemann, J. P., Tanner, E. V., Hiederer, R., and Kapos, V. (2014). Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Management* 5 (1): 81–91.

Schlesinger, W. H. and Andrews, J. A. (2000). Soil respiration and the global carbon cycle. *Biogeochemistry* 48 (1): 7–20.

Silver, W., Ostertag, R., and Lugo, A. (2000). The potential for carbon sequestration through reforestation of abandoned tropical agricultural and pasture lands. *Restoration ecology* 8 (4): 394–407.

Stockmann, U., Adams, M. A., Crawford, J. W., Field, D. J., Henakaarchchi, N., Jenkins, M., Minasny, B., McBratney, A. B., de De Courcelles, V. R., Singh, K., *et al.*, (2013). The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agriculture, Ecosystems & Environment* 164: 80–99.

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., Hong, S. Y., Rawlins, B. G., and Field, D. J. (2015). Global soil organic carbon assessment. *Global Food Security* 6: 9–16.

Tiessen, H, Cuevas, E, and Chacon, P. (1994). The role of soil organic matter in sustaining soil fertility. *Nature* 371 (6500): 783.

Van Gestel, N., Shi, Z., Van Groenigen, K. J., Osenberg, C. W., Andresen, L. C., Dukes, J. S., Hovenden, M. J., Luo, Y., Michelsen, A., Pendall, E., *et al.*, (2018). Predicting soil carbon loss with warming. *Nature* 554 (7693): E4.

Van Veen, J. and Paul, E. (2011). Organic carbon dynamics in grassland soils. 1. Background information and computer simulation. *Canadian Journal of Soil Science*.

Watson, S. J., Luck, G. W., Spooner, P. G., and Watson, D. M. (2014). Land-use change: incorporating the frequency, sequence, time span, and magnitude of changes into ecological research. *Frontiers in Ecology and the Environment* 12 (4): 241–249.

Yin, X., Goudriaan, J., Lantinga, E. A., Vos, J., and Spiertz, H. J. (2003). A flexible sigmoid function of determinate growth. *Annals of botany* 91 (3): 361–371.

Zhang, H. K. and Roy, D. P. (2017). Using the 500 m MODIS land cover product to derive a consistent continental scale 30 m Landsat land cover classification. *Remote sensing of environment* 197: 15–34.

Zou, X and Bashkin, M (1998). Soil carbon accretion and earthworm recovery following revegetation in abandoned sugarcane fields. *Soil Biology and Biochemistry* 30 (6): 825–830.

Chapter 8

General discussion, conclusions and future research

8.1 General discussion

8.1.1 Machine learning

A common theme running through this thesis is the use of advanced machine learning (ML) methods such as neural networks (NN). Thanks to their flexible structure and training options, NN can be used in a diverse number of situations. If we consider the input data, in Chapters 4 and 5 we used spectral data encoded as a 2D spectrogram to train a model, although spectral data can also be processed directly as a 1D array (Ng *et al.*, 2019). In Chapter 2, we used information derived from a stack of rasters (3D) as input for the NN. NNs also have the capability to process 4D input data (Yang *et al.*, 2015), opening the possibility to use them in space-time applications. Particularly with complex data structures (e.g. 1D sequence of wavelets, 2D spectrogram of 3D stack of covariates), the use of advanced NNs such as a convolutional NN (CNN) gives the model the capacity to preserve the spatial structure of the data. Other methods such as tree-like models (Cubist, random forest) are not designed to handle complex input data which should always be converted to a 1D vector, hence losing its spatial structure. In terms of model outputs, in this thesis we only applied NNs to predict continuous outputs but NNs perform very well in classification tasks (Krizhevsky *et al.*, 2012). The way we can train NNs is also quite unique. In this work, we trained NN

using different fractions of the data at a time (online learning; Chapter 3), dissected a NN to transfer the knowledge for one model to another (transfer learning; Chapter 5), and simultaneously predict multiple properties in a synergistic approach (multi-task learning; Chapter 2 and Chapter 4).

Multi-task learning

By using multi-task learning, NNs have the capability to simultaneously predict or classify multiple outputs. We applied this training technique in the context of digital soil mapping (DSM; Chapter 2) and soil spectroscopy (Chapter 4) to simultaneously predict multiple soil depths and properties, respectively. Multi-task learning was able to provide better predictions than using individual models by sharing common features between tasks. The mechanisms on how multi-task learning works are not completely clear. Some authors suggest that the performance depends on how correlated the tasks are (Xu *et al.*, 2017), but that does not explain why pH showed a degraded performance (Chapter 4) even when its correlation with some properties was relatively high in that dataset (e.g. 0.5 with clay). Ma *et al.*, (2015) concluded that a single-task model performs better for large datasets, while for the smaller datasets a multi-task model performs better. That is the opposite to the trend that we observed in Chapter 4. We still need to study the behaviour of these models in other scenarios, but we think it is useful to consider the complexity of the overall task. For instance, a really complex model to simultaneously predict multiple depths and properties for sure needs many observations to untangle all the relationships and it will overfit a small dataset. At this stage, I do recommend the use of multi-task learning to simultaneously predict multiple soil properties (or depths) but always comparing it with single-tasks models. Another recommendation is to be judicious when selecting the soil properties to predict. Many combinations could lead to spurious relationships and wrong interpretations (Zhou *et al.*, 2017), so properties that are well-known to be related should be prioritised.

New good practices

Thanks to the effort of some groups to rescue soil legacy data (Arrouays *et al.*, 2017), and cheaper and faster methods to analyse soil samples, there is more soil data available than ever before. This data availability not only allows us to use new ML algorithms

which usually require more observations but also opens the door to new ways to train those models. As shown in Chapter 1, soil scientists have been using ML for many years and there is a clear tendency towards their higher adoption.

An important part of model development is validation. Literature traditionally recommends that an independent, unseen (by the model) dataset should be used as validation (Kohavi, 1995). In practice, the data is usually partitioned into training and validation datasets. A more stable solution is the use of k-cross-validation where the dataset is partitioned into k groups, where $k-1$ groups are used for training and 1 group for validation, repeating the training k times, each with a different validation group. When data availability is a limitation, researchers resort to techniques such as n-cross-validation or "leave-one-out" validation to make the most of the available data (Stevens *et al.*, 2008; Pasini, 2015).

A new generation of models based on neural networks (NNs) have been introduced in the later years, which have revolutionised many fields. Deep learning (DL) models, consisting of multiple hidden layers of neurons, have many parameters (from hundreds to millions) which need to be fitted in the training process. This is the reason why they usually need access to large sample sizes. A second characteristic of these models is that they have a considerable number of hyper-parameters. Hyper-parameters are parameters that are not learned from the data during the training phase and include number of iterations during the training, learning rate, layer parameters, number of layers, etc. A common practice when training DL models is to split the original dataset into 3 sub datasets: training, validation and test. The training dataset is used to learn the parameters, the validation dataset to compare models fitted with different hyper-parameters in order to find the optimal combination, and the test dataset as the independent, unseen data.

In soil sciences, ML algorithms are usually trained using the traditional train/validation split or cross-validation (Keskin *et al.*, 2019; Liang *et al.*, 2019), or even no validation (Feng *et al.*, 2019), except for some studies based on DL or with engineering background (e.g. Reale *et al.*, 2018), including the publications derived from Chapters 2-5, which use a train/validation/test split. Considering the increasing size of datasets, I think soil scientists should transition towards the implementation of some DL practices such as dataset split and hyper-parameter optimisation (Bergstra and Bengio, 2012; Snoek *et al.*, 2012), not only for NNs but for any algorithm that

has hyper-parameters. Some potential candidates usually used in soil literature (see Chapter 1) are random forest, cubist, classification and regression trees, and support vector machines. Most of the implementations of these algorithms have sensible default hyper-parameters but some studies report an important impact of them in their results (Mutanga *et al.*, 2012; Lu *et al.*, 2018).

Commercial applications

This work explores the use of ML in soil sciences, and we have shown applications that can aid in different research aspects including generation of better predictive models, building models collaboratively or exploring the current literature. But ML has extended beyond research and companies are very welcoming to this technology, especially in applications such as computer vision, speech recognition, natural language processing, and robot control (Jordan and Mitchell, 2015).

It is not hard to imagine a commercial application of approaches such as soil properties prediction using vis-NIR spectroscopy (Chapter 4), either in the laboratory or the field. While in research there are some transparency requirements, including describing the methods and data used, companies are usually very secretive about their methods since they are a trade secret that gives them a competitive advantage. Considering that lack of transparency, how can we be sure that the predictions of their models are good? There is not a unique answer but it should include at least some accuracy and uncertainty assessment, and information about the range of soils used during training.

Uncertainty assessment is an important requirement for any model, especially if the predictions are going to be used to guide decision-making. Potential approaches to generate prediction intervals include the use of bootstrap when training the model, as shown in Chapter 4, effectively making predictions with many models trained with subsets of the original data; approaches such as fuzzy k-means with extragrades (Tranter *et al.*, 2010), which defines areas within the covariate space, with different levels of uncertainty, where the new observation (to be predicted) can be placed; and the use of Bayesian optimisation approaches (Gal and Ghahramani, 2016; Snoek *et al.*, 2015).

In terms of reporting soil type coverage, different approaches can be applied. A

simple, perhaps over-confident method can be reporting the geographical extent from where the soil samples used during training were collected (e.g. Tomasella *et al.*, (2000) and Børgesen and Schaap (2005)), or a broad soil classification based in the soil characteristics such as "sandy soils" (e.g Schaap and Bouten (1996) and Shaw *et al.*, (2000)). A better approach, based on the covariate space of the samples used during training is fuzzy k-means with extragrades, which has the benefit of describing both, coverage and uncertainty levels.

Even if uncertainty levels and coverage are reported, another factor to consider is how much we should trust in companies and their reports. Especially for applications involving public funding, but generally as a consumer protection measure, this type of products should be certifiable, in the same way many soil laboratories are. A usual approach is the use of reference materials (Dybczyński *et al.*, 1979; Pueyo *et al.*, 2001; Ahmed *et al.*, 2017), which should be consistent with the model coverage reported. The properties measured in the reference materials should fall within the prediction interval produced by the model, with a confidence defined for each application.

8.1.2 Using new (old) information

On the use of field descriptive soil data

The information contained in soil surveys is extremely valuable since it represents the mental model of the soil scientists that performed the task. Bui (2004) defines a soil map as a structured representation of knowledge about soils' spatial distribution. Since maps are usually created from point information, the description of each point contains enough information to draw conclusions about a) the individual site and b) the relationship between all the sites (assuming we have information about their location).

Even if information from soil surveys can be collected in a structured database with the aid of handheld devices, the general case is that a large amount of information is captured as qualitative descriptions. This information is usually hard to process and integrate with traditional data analysis workflows since it is in the form of natural language, and it is usually discarded in electronic databases. In Chapter 6 we make use of a special category of ML which is capable of capturing the high level structures contained in a group of documents and assigning each work a position within a high dimensional vector space.

We see a great potential on using textual data to complement numerical, analytical data. We presented one example of the use of these word embeddings (vector representation of words) where we emulated a workflow for numerical data in the context of a soil numerical classification system but the future applications of these domain-specific embeddings are vast. For instance, in a derived study, Fuentes and Padarian (2019) used the word embeddings to perform a 3D lithological mapping based on bore descriptions. Using the embeddings allowed us to use advanced workflows usually reserved for numerical data, leading, for example, to a better assessment of the uncertainty of the generated maps.

8.1.3 Collaboration and data sharing

The cornerstone of collaboration is trust. Arguably, the usual way of sharing information is that most of the parties send some files over the internet while one party acts as a collator. Depending on the terms of the collaboration, a series of promises are in place, which might include receiving a copy of the collated version of the dataset, not using the data for other studies, etc. If any misbehaviour is suspected, there is a breach of trust and the chances of the collaboration succeeding are reduced. In order to prevent misbehaviour from some of the parties, there are alternatives that minimise or eliminate this need for trust.

In Chapter 3, we introduced the use of online learning as a method to train models using datasets held by different parties while preserving privacy. This is a fully opaque approach, where the data is never transferred between parties but, instead, the model is trained while it "travels" from one dataset to another. The results of this thesis show that the performance of the model is not significantly affected compared to a model trained using all the data. This could be a potential solution when countries can not share their complete soil databases due to legal restraints.

The opposite situation is when all the parties are willing to share their data towards a common project. Given that the general approach is to centralise the data into a single-source, the risk of a party or dataholder misbehaving still exists. A potential solution is the use of a distributed ledger (blockchain). In simple terms, a blockchain is a linked sequence of records of the transactions of digital assets. A blockchain has various characteristics that makes it an interesting candidate for an intra- or inter-institutional

database. It is decentralised, since each node keeps a copy of the blockchain; it is immutable, since it is not possible to remove or edit a transaction; and each party is responsible for their own data, since each transaction is cryptographically signed by the owner.

If data privacy is a concern, we recommend sharing models instead of the data used to train them. The model should include metadata on how it was trained and some summary information about the data used. FAO, a global organisation, promotes global collaborations such as Global Soil Information System (GLOSIS) and Global Soil Laboratory Network (GLOSOILAN) for information sharing. As data trust becomes an issue, technology such as blockchain should be considered. As far as we are aware, there are no ready-to-use implementations that aid the process for online learning (or a blockchain soil database) but we have plans of developing such a platform in the near future.

8.1.4 Global mapping

During this thesis we improved on many methodological aspects of soil modelling. The methods may seem like disjointed contributions but they actually fit within a single modelling framework (Fig. 8.1) in order to solve more complex challenges such as soil mapping at the global scale. Global and local models, either in the context of DSM or soil spectroscopy, can be interconnected in a continuous feedback privately sharing local knowledge to build global models via online learning and applying the general knowledge of those global models to improve predictions at the local scale via transfer learning. The data derived from spectral models can be also integrated in a multi-task CNN model as part of the local datasets (Wadoux *et al.*, 2019). Potentially, information derived from descriptive data via language models (word embeddings) can complement numerical data. In Fuentes and Padarian (2019) we demonstrated the use of word embeddings to perform a 3D lithological mapping and that could be also implemented within a soil mapping framework.

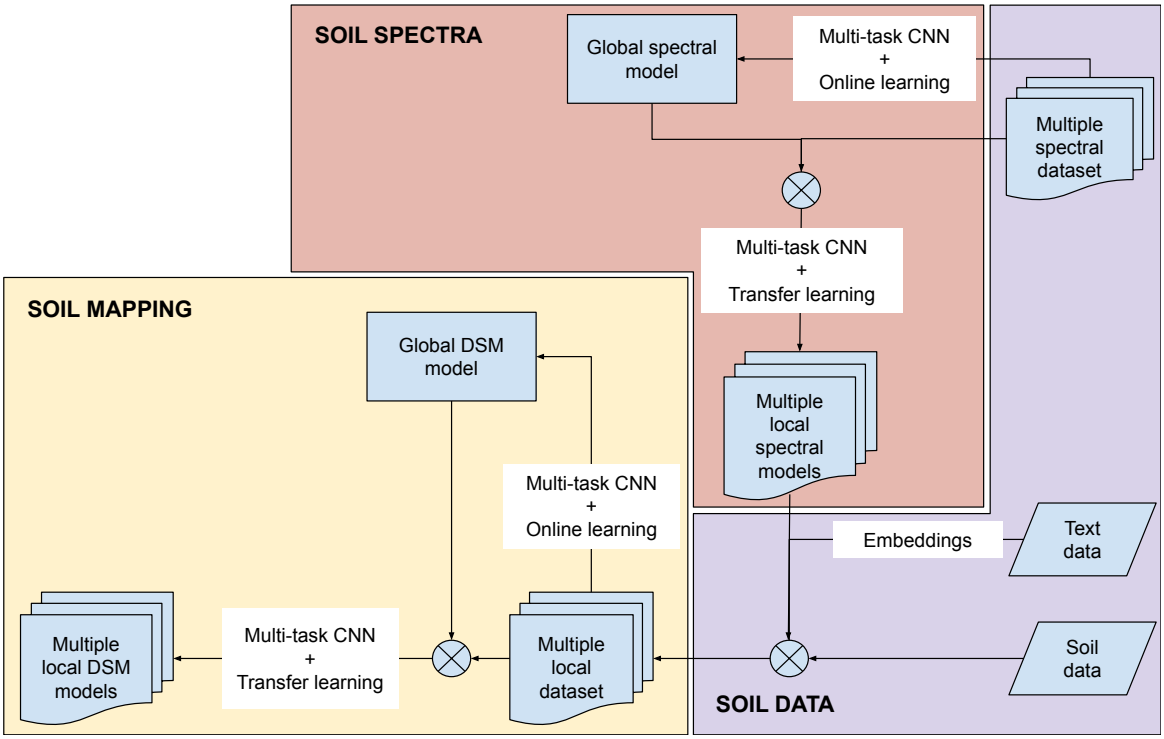


Fig. 8.1: Integration of the methods developed in this thesis into a single modelling framework. CNN: convolutional neural network.

8.2 Overall research conclusions

The research presented here has been successful at applying the latest advances in machine learning (ML) to improve upon current approaches for soil modelling and creating new opportunities to utilise data that has been generally disregarded.

In the context of spatial modelling, the use of a multi-task CNN brings an ML model closer to the mental model of a soil scientist by including two important components, namely contextual information surrounding a soil observation and information from other properties by predicting them simultaneously. In this research, "other properties" correspond to other soil depth intervals, effectively improving prediction of deeper layers which are poorly represented by the covariates used within a DSM framework.

We expanded the multi-task learning approach to predict multiple properties in the context of soil spectroscopy. The use of a multi-task CNN model to predict multiple soil properties from soil spectral data yielded a dramatic reduction in the

error compared with traditional methods. In addition, the proposed approach does not require pre-processing of the data. An important factor to consider is that, in order to obtain a good model, a relatively large soil spectral dataset is required.

Large spectral databases are not easy to generate and they are usually products of operations to characterise large extents. The model generated from large, general datasets usually perform poorly when applied in a different context such as local scale. We successfully used an advanced model building technique (i.e. transfer learning) to use the general knowledge of a large dataset to significantly improve local models, effectively connecting both scales.

One of the options to generate large soil datasets is collaboration between different parties but it is usually hindered by concerns such as privacy and confidentiality. We implemented an advanced model training technique (online learning) to replace the traditional workflow where data needs to be sent to a centralised party. Instead, the model can be trained on a single sub-group of observations at the time. The workflow then is inverted and the model is moved between parties, not the data. This method proves to be as effective as training the model in a centralised location with access to all the information.

Most of the data utilised for soil modelling is numerical. However, there is valuable information contained on descriptive, textual data that have not been used. We built a language model based on natural language processing to represent textual data as vectors, capturing syntactic and semantic relationships between their constituting words. By only using the word embedding to represent soil profile descriptions we were able to emulate an analysis performed with numerical data obtained in the laboratory, and obtain similar results.

At the moment, some challenges from the soil science community are not solvable by directly applying ML methods, mainly due to data limitations. In such cases, using ML models as part of a semi-mechanistic approach is recommended. We successfully applied this approach to build a complex space-time model to assess soil organic carbon (SOC) at the global scale. First, a machine learning model was used to link environmental factors (proxies for soil formation factors) and soil profile observations. Then, by using SOC dynamics and a landcover tracking component based on remote sensing, the SOC contents were propagated through time at one-year steps.

Machine learning methods have been embraced by the soil community and their

adoption is increasing. In the particular case of neural networks, their flexibility in terms of structure and training makes them a good candidate to improve on current soil modelling approaches. Since they are usually considered "black-box" models, a main focus of future research should be on improving their interpretability.

8.3 Future work

There are many opportunities for future work and some of these have been briefly mentioned in previous chapters. Opportunities include:

(i) *Transfer learning in the context of DSM:* The method presented in Chapter 5 is a general approach to transfer learning from a "general" to a "local" model. This was demonstrated in the context of predicting soil properties from soil spectral data. A logical step forward is to evaluate the effectiveness of this approach using different data. There are a few initiatives promoting the development of soil maps at global extent (FAO, GlobalSoilMap), and also projects generating soil maps at continental scale (maps derived from LUCAS dataset in Europe; TERN in Australia). All those spatial models contain valuable information that can be utilised (transferred) to local models, either for regional or field applications.

(ii) *Transfer learning to aid decision-making at field scale:* Related with the previous point, transfer learning, either in DSM or spectroscopy, can also be applied at field scale. Potentially, this can reduce the burden of a field scale campaign in terms of the number of samples required to properly capture the local soil variation. It is pertinent to study if a global map is directly transferable to the field scale or if an intermediate, national or regional step is necessary.

(iii) *Use of text description:* The word embeddings introduced in Chapter 6 have the potential to be included in many downstream applications. It is necessary to evaluate if it is possible to improve current workflows by including these embeddings, especially when numerical data is not available or it is uncertain.

(iv) *Model interpretability:* Considering that in this thesis we introduced the applications of deep learning models in soil science, specifically convolutional neural networks, there is a long way to go from the increased predicting capacity

to actually extract knowledge from these models and improve our understanding of soils. This is achieved by implementing methods to interpret the internal structure of the neural networks, and we foresee a large number of studies focusing on this topic in the following years.

(v) *Global SOC mapping:* We have seen some platform for "real-time" monitoring of surface water (Donchyts *et al.*, 2016) and forest (Hansen *et al.*, 2013) at the global scale, and we think a similar platform can be implemented for SOC. At the moment, the model is restricted by the 1-year temporal resolution of MODIS landcover data but it can be adapted for more frequent images as they become available.

8.4 References

Ahmed, O., Habbani, F. I., Mustafa, A., Mohamed, E., Salih, A., and Seedig, F. (2017). Quality assessment statistic evaluation of X-ray fluorescence via NIST and IAEA standard reference materials. *World Journal of Nuclear Science and Technology* 7 (02): 121.

Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., *et al.*, (2017). Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ* 14: 1–19.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13 (Feb): 281–305.

Børgesen, C. D. and Schaap, M. G. (2005). Point and parameter pedotransfer functions for water retention predictions for Danish soils. *Geoderma* 127 (1-2): 154–167.

Bui, E. N. (2004). Soil survey as a knowledge system. *Geoderma* 120 (1-2): 17–26.

Donchyts, G., Baart, F., Winsemius, H., Gorelick, N., Kwadijk, J., and Van De Giesen, N. (2016). Earth's surface water change over the past 30 years. *Nature Climate Change* 6 (9): 810.

Dybczyński, R., Tugsavul, A, and Suschny, O. (1979). Soil-5, a new IAEA certified reference material for trace element determinations. *Geostandards Newsletter* 3 (1): 61–87.

Feng, Y., Cui, N., Hao, W., Gao, L., and Gong, D. (2019). Estimation of soil temperature from meteorological data using different machine learning models. *Geoderma* 338: 67–77.

Fuentes, I. and Padarian, J. (2019). 3D lithological mapping of borehole descriptions using word embeddings. *Computers & Geosciences*. Under review.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *International conference on machine learning*: pp. 1050–1059.

Hansen, M., Potapov, P., Moore, R, Hancher, M, Turubanova, S., Tyukavina, A, Thau, D, Stehman, S., Goetz, S., Loveland, T., *et al.*, (2013). High-resolution global maps of 21st-century forest cover change. *Science* 342 (6160): 850–853.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science* 349 (6245): 255–260.

Keskin, H., Grunwald, S., and Harris, W. G. (2019). Digital mapping of soil carbon fractions with machine learning. *Geoderma* 339: 40–58.

Kohavi, R. *et al.*, (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*. Vol. 14. 2. Montreal, Canada: pp. 1137–1145.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*: pp. 1097–1105.

Liang, Z., Chen, S., Yang, Y., Zhao, R., Shi, Z., and Rossel, R. A. V. (2019). National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. *Geoderma* 335: 47–56.

Lu, W., Lu, D., Wang, G., Wu, J., Huang, J., and Li, G. (2018). Examining soil organic carbon distribution and dynamic change in a hickory plantation region with Landsat and ancillary data. *Catena* 165: 576–589.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling* 55 (2): 263–274.

Mutanga, O., Adam, E., and Cho, M. A. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation* 18: 399–406.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., and McBratney, A. B. (2019). Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* 352: 251–267.

Pasini, A. (2015). Artificial neural networks for small dataset analysis. *Journal of thoracic disease* 7 (5): 953.

Pueyo, M, Rauret, G, Bacon, J., Gomez, A, Muntau, H, Quevauviller, P., and López-Sánchez, J. (2001). A new organic-rich soil reference material certified for its EDTA-and acetic acid-extractable contents of Cd, Cr, Cu, Ni, Pb and Zn, following

collaboratively tested and harmonised procedures. *Journal of Environmental Monitoring* 3 (2): 238–242.

Reale, C., Gavin, K., Librić, L., and Jurić-Kaćunić, D. (2018). Automatic classification of fine-grained soils using CPT measurements and Artificial Neural Networks. *Advanced Engineering Informatics* 36: 207–215.

Schaap, M. G. and Bouten, W. (1996). Modeling water retention curves of sandy soils using neural networks. *Water Resources Research* 32 (10): 3033–3040.

Shaw, J., West, L., Radcliffe, D., and Bosch, D. (2000). Preferential flow and pedotransfer functions for transport properties in sandy Kandiudults. *Soil Science Society of America Journal* 64 (2): 670–678.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In: *Advances in neural information processing systems*: pp. 2951–2959.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. (2015). Scalable bayesian optimization using deep neural networks. In: *International conference on machine learning*: pp. 2171–2180.

Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., and Ben-Dor, E. (2008). Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma* 144 (1-2): 395–404.

Tomasella, J., Hodnett, M. G., and Rossato, L. (2000). Pedotransfer Functions for the Estimation of Soil Water Retention in Brazilian Soils. *Soil Science Society of America Journal* 64: 327.

Tranter, G, Minasny, B., and McBratney, A. (2010). Estimating Pedotransfer Function Prediction Limits Using Fuzzy k-Means with Extragrades. *Soil Sci. Soc. Am. J.* 74 (6): 1967–1975.

Wadoux, A. M.-C., Padarian, J., and Minasny, B. (2019). Multi-source data integration for soil mapping using deep learning. *Soil* 5 (1): 107–119.

Xu, Y., Ma, J., Liaw, A., Sheridan, R. P., and Svetnik, V. (2017). Demystifying multitask deep neural networks for quantitative structure–activity relationships. *Journal of chemical information and modeling* 57 (10): 2490–2504.

Yang, J., Nguyen, M. N., San, P. P., Li, X. L., and Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human

activity recognition. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Zhou, L., Pan, S., Wang, J., and Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing* 237: 350–361.