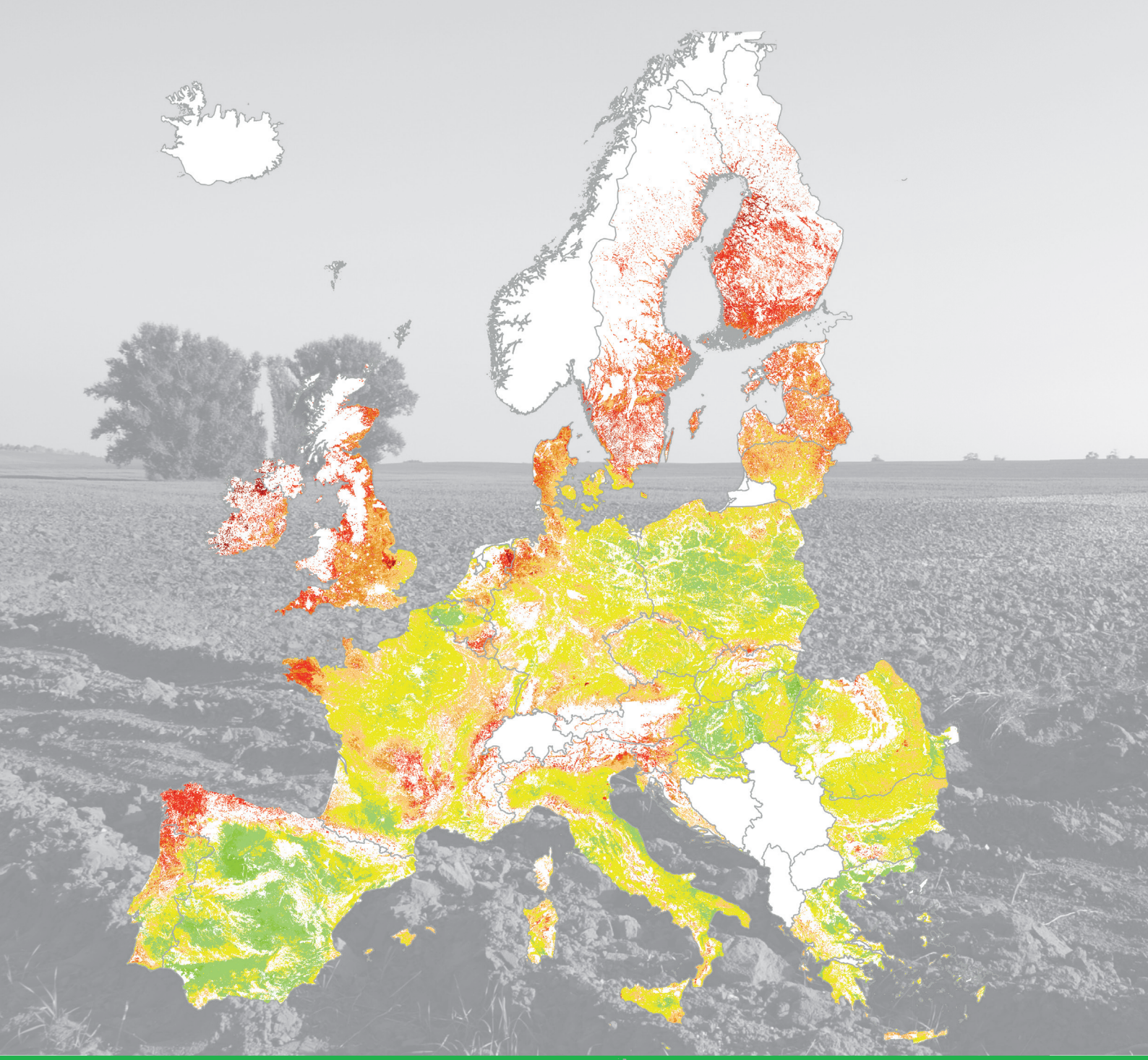

A soil organic matter map for arable land in the EU

L. de Sousa, F. van den Berg and G.B.M. Heuvelink

WAGENINGEN
UNIVERSITY & RESEARCH

ISRIC
World Soil Information

A soil organic matter map for arable land in the EU

L. de Sousa[1], F. van den Berg[2] and G.B.M. Heuvelink[1]

1 ISRIC

2 Wageningen Environmental Research

This research was subsidised by the Dutch Ministry of Agriculture, Nature and Food Quality (project number BO-43-102.01-004).

Wageningen Environmental Research
Wageningen, February 2022

Reviewed by:
Martin Knotters, research scientist (Wageningen Environmental Research)

Approved for publication:
dr. ir. J.A. de Vos

Report 3126
ISSN 1566-7197

WAGENINGEN
UNIVERSITY & RESEARCH

ISRIC
World Soil Information

De Sousa, L., F. van den Berg and G.B.M. Heuvelink, 2022. *A soil organic matter map for arable land in the EU.* Wageningen, Wageningen Environmental Research, Report 3126. 46 pp.; 26 fig.;11 tab.; 31 ref.

For the assessment of the leaching of crop protection products to groundwater according to EC regulation 1107/2009 a tiered approach has been adopted. A spatially-distributed leaching model is used in tier 3b and tier 4 assessments as described in the EU Guidance. One of the key factors that influences the leaching to groundwater is the soil organic matter content. This content not only depends on factors such as climate and terrain morphology, but also on land use. As the current EFSA organic matter map does not take land use into account, a study was done to investigate whether ISRIC's SoilGrids product could be used to improve the soil organic matter map for arable land in Europe. Machine learning algorithms were applied using the available data in the SoilGrids database. A 10-fold cross-validation procedure was used to assess the accuracy of the predictions. The best results were obtained using an ensemble model whilst training the models with data on the log-scale and limiting the observations to organic matter contents below 5%, which is common in soils in arable land. As expected, the predictions made by SoilGrids were substantially lower and more accurate than those from the EFSA soil organic matter map.

Keywords: Cross-validation, EFSA map, Ensemble, GeoPEARL, Gradient Boosting, Machine learning, Random Forest, SoilGrids.

The pdf file is free of charge and can be downloaded at https://doi.org/10.18174/556312 or via the website www.wur.nl/environmental-research (scroll down to Publications – Wageningen Environmental Research reports). Wageningen Environmental Research does not deliver printed versions of the Wageningen Environmental Research reports.

© 2022 Wageningen Environmental Research (an institute under the auspices of the Stichting Wageningen Research), P.O. Box 47, 6700 AA Wageningen, The Netherlands,
T +31 (0)317 48 07 00, www.wur.nl/environmental-research. Wageningen Environmental Research is part of Wageningen University & Research.

- Acquisition, duplication and transmission of this publication is permitted with clear acknowledgement of the source.
- Acquisition, duplication and transmission is not permitted for commercial purposes and/or monetary gain.
- Acquisition, duplication and transmission is not permitted of any parts of this publication for which the copyrights clearly rest with other parties and/or are reserved.

Wageningen Environmental Research assumes no liability for any losses resulting from the use of the research results or recommendations in this report.

In 2003 Wageningen Environmental Research implemented the ISO 9001 certified quality management system.
Since 2006 Wageningen Environmental Research has been working with the ISO 14001 certified environmental care system.
By implementing the ISO 26000 guideline, Wageningen Environmental Research can manage and deliver its social responsibility.

Wageningen Environmental Research report 3126 | ISSN 1566-7197

Photo cover: Soil organic matter map for arable land in the EU (this report)

Contents

Verification

Report: 3126
Project number: 5200046712

Wageningen Environmental Research (WENR) values the quality of our end products greatly. A review of the reports on scientific quality by a reviewer is a standard part of our quality policy.

Approved reviewer who stated the appraisal,

position: Research scientist

name: dr.ing. M Knotters

date: 25 October 2021

Approved team leader responsible for the contents,

name: dr.ir. J.A. de Vos

date: 12 November 2021

Preface

One of the key factors that influence the leaching to groundwater is the soil organic matter content. Furthermore, the soil organic matter content depends on the land use. In the current EFSA dataset the effect of land use has not been taken into account in the preparation of the soil organic matter map for the EU. This shortcoming could be remedied using the SoilGrids model. This model has been developed to estimate the soil organic carbon content based on soil organic carbon measurements and correlated environmental variables. Therefore a study was set up to investigate the use of the SoilGrids model to generate a soil organic matter map for arable land in Europe. For this research, a collaboration project was started by Wageningen Environmental Research (WENR) and the International Soil Reference and Information Centre (ISRIC). The current report describes the procedure to generate a new soil organic matter map for Europe, the resulting maps and comparisons with available data and the EFSA map.

Summary

In the EU the leaching of crop protection products to groundwater is assessed according to EC regulation 1107/2009. For the assessment of the leaching potential in the first tier, the FOCUS Groundwater scenarios have been developed. These scenarios are considered to represent significant areas of Europe. In tier 3b and tier 4 assessments as described in the EU Guidance on the assessment of the risk of leaching to groundwater, a spatially-distributed leaching model is used. For such models spatial data are needed on soil physical type and soil properties, climate, hydrotype and land use.

One of the key factors that influence the leaching to groundwater is the soil organic matter content. The soil organic matter content depends also on land use, which is not taken into account in the EFSA 2012 dataset with GIS data on the soil organic matter for the EU. When using the EFSA dataset, the soil organic matter in soils with arable land use is over-estimated. This study aimed to account for the effect of land use on soil organic matter and obtain an improved soil organic matter map for Europe.

An improved soil organic matter map for Europe could be prepared using SoilGrids 2017. This model for digital soil mapping as developed by ISRIC – World Soil Information predicts inter alia the soil organic matter content based on soil organic carbon measurements and correlated environmental variables using 'machine-learning'. In this study, a pilot study was done first to assess the accuracy of the existing SoilGrids 2017 soil organic matter maps for the Netherlands and the EU. The SoilGrids 2017 soil organic matter map clipped to the Netherlands was compared with a soil organic matter map that has been prepared to be used for assessments of the leaching to groundwater in the Dutch authorization procedure for plant protection products (GeoPEARL). The soil organic matter map for the Netherlands showed that the contents predicted by SoilGrids 2017 were somewhat higher. The most likely explanation for this difference is that the map for the soil organic matter for GeoPEARL had been obtained by the log-transformation of the soil organic matter contents prior to model calibration. Using that approach for soils with comparatively low organic matter contents, i.e. less than 5%, often to be observed in areas with arable land use, low organic matter contents are given more weight in the calibration procedure. The SoilGrids 2017 soil organic matter map for Europe was also compared with the soil organic matter map prepared on the basis of the EFSA dataset 2012. When comparing with the LUCAS dataset, the SoilGrids 2017 model over-predicted somewhat the soil organic matter content, but this over-prediction was much smaller than the over-prediction of the EFSA dataset.

To further improve the soil organic matter map using the SoilGrids model a tailored approach was adopted by applying several modifications to the calibration strategy. Firstly, the area for which predictions were made was limited to the area with arable land use as defined in the Corine Land Cover 2012 map. Furthermore, the SoilGrids model was calibrated with soil organic carbon observations from arable land only. As SoilGrids predicts the organic carbon content, a conversion factor was applied to express the result as the content of organic matter. The tailor-made SoilGrids map of the predicted organic matter content for soils in arable land in Europe was identified as SoilGrids 2018. This map was prepared on the basis of more than 40 000 soil organic carbon observations and 200 maps on environmental variables from different domains, such as climate, ecology, geology, land use, terrain morphology, vegetation and water. Soil organic matter maps were generated by the machine learning algorithms Random Forest and Gradient Boosting, and an Ensemble model that produces a weighted-average of the outcome of the Random Forest and Gradient Boosting models. Maps for the organic matter content were made for four soil depths, i.e. 0.15 m, 0.45 m, 0.8 m and 1.2 m below the soil surface. The results show that the predicted organic matter content declines strongly with increasing depth, with the largest decrease occurring between 0.15 and 0.45 m depth.

For the assessment of the risk of leaching to groundwater, it is important to predict the soil organic matter content in the range of a few percent (2 - 4%) as accurately as possible. Therefore, similar to the development of the new soil organic matter map for GeoPEARL, the soil organic matter contents

were first transformed to values on a log-scale before processing them by SoilGrids. The accuracy of the resulting SoilGrids 2018 predictions was assessed using a 10-fold cross-validation procedure. Using this procedure, the observations are split into ten different subsets of similar size where nine subsets are used for model calibration and one subset for validation. The Mean Error for the RF and GB models were calculated to be 0.58% and 0.43% of organic matter, respectively. The Root Mean Squared Error (RMSE) values ranged from 4.59% to 5.08%. For the subset of soil organic matter data below 10%, the Mean Error for the RF and GB models was calculated to be much smaller, 0.09% and 0.01%, respectively. The associated RMSE values were calculated to be 1.32% and 1.55% for the RF and GB models, respectively. By calibration with data on the log scale, the accuracy of the predicted organic matter content for the subset of data below 10% soil organic matter was much better. By limiting the data to those with organic matter contents below 5% instead of 10% there was only a slightly better performance of the model.

The predictions obtained with the Random Forest model were more accurate than those obtained with the Gradient Boosting model. The results of the Ensemble model were slightly more accurate than the results obtained with the Random Forest model and this model was subsequently used to generate the soil organic matter map for soils in arable land in Europe. The results of the cross-validation procedure showed that with the calibration strategy applied in SoilGrids 2018 sufficiently accurate predictions could be made for the organic matter content in these soils.

Based on the results of the Ensemble model calibrated with organic carbon contents on the log-scale, soil organic matter maps were prepared for the EU at four depths, at 0.15 m, 0.45 m, 0.80 m and 1.2 m. The map generated for a depth of 0.15 m was compared with the EFSA 2012 map. In general, the predicted value using SoilGrids 2018 was substantially lower than the organic matter content according to the EFSA map, in particular for Central Europe. This difference can be explained by the fact that the EFSA map considers organic matter data for all types of land use, whereas the SoilGrids 2018 results concern only areas with arable land use.

The soil organic matter maps for the EU using SoilGrids 2018 were also compared with those prepared using SoilGrids 2017. The SoilGrids 2018 predictions are generally lower than the SoilGrids 2017 predictions, due to the calibration with log-transformed soil organic matter contents and considering observations in arable land only.

The results of the validation using the LUCAS dataset showed that SoilGrids 2017 and in particular EFSA 2012 overestimated the organic matter content in soils in arable land, whereas this was not the case with SoilGrids 2018.

The results of this study show that an improved soil organic matter map for the EU can be made using SoilGrids 2018 in comparison with the EFSA 2012 and SoilGrids 2017 maps. The SoilGrids 2018 map obtained in this study could be used as input for a soil schematisation for a spatially-distributed model to assess the leaching of pesticides to groundwater at the EU level. Such a schematisation is currently being developed by the SETAC Working Group on Spatially Distributed Leaching Modelling (SETAC-SDLM).

SoilGrids 2018 could be improved further by adding new data on organic matter content in soils for the calibration procedure. In doing so, it is important to pay attention to the spatial distribution of the observations. The improvement could be important for areas with a relatively limited number of observations. In addition, the set of environmental variables to be used for the prediction could be extended and improved.

Samenvatting

In de EU wordt de uitspoeling van gewasbeschermingsmiddelen naar het grondwater beoordeeld op basis van EC Verordening 1107/2009. Voor de beoordeling van het risico op uitspoeling zijn FOCUS Grondwater scenario's ontwikkeld. Deze scenario's worden representatief geacht voor landbouwgebieden in Europa. In tier 3a en tier 4 van de beoordeling wordt een ruimtelijk-gedistribueerd model gebruikt. Voor een dergelijk model zijn ruimtelijke gegevens nodig over het fysische bodemtype en de bodemeigenschappen, klimaat, hydrotype en landgebruik.

Een van de belangrijkste factoren die de uitspoeling van gewasbeschermingsmiddelen naar het grondwater beïnvloeden is het organische-stofgehalte van de bodem. Het organische-stofgehalte hangt ook af van het landgebruik, waarmee in de EFSA 2012 dataset met GIS data voor het organisch-stofgehalte in Europa geen rekening is gehouden. Dat resulteert bij gebruik van de EFSA dataset in een overschatting van het organische-stofgehalte in akkerland. Deze studie had tot doel wel rekening te houden met het effect van het landgebruik op het organische-stofgehalte om daarmee een verbeterde kaart van het organische stof gehalte voor Europa te verkrijgen.

Een betere kaart van het organische-stofgehalte voor Europa zou verkregen kunnen worden met SoilGrids 2017. Dit door ISRIC – World Soil Information ontwikkelde model voor digitale bodemkartering voorspelt onder andere het organisch-stofgehalte in de bodem op basis van metingen van het organische-koolstofgehalte en metingen van gecorreleerde omgevingsvariabelen met behulp van 'machine learning'. In deze studie werd eerst een pilot studie uitgevoerd om de nauwkeurigheid van de bestaande SoilGrids 2017 organische-stof kaarten voor Nederland en de EU te evalueren. De uitsnede van de kaart van SoilGrids 2017 voor Nederland werd daartoe vergeleken met de organische-stof kaart die ontwikkeld is voor de beoordeling van de uitspoeling naar het grondwater volgens de procedure voor de toelating van gewasbeschermingsmiddelen in Nederland (GeoPEARL). Deze vergelijking bracht aan het licht dat de door SoilGrids 2017 voorspelde gehalten wat hoger waren. De meest waarschijnlijke verklaring hiervoor is dat de kaart voor het organische-stofgehalte in de bodem voor GeoPEARL was verkregen na log-transformatie van het organische-stofgehalte voorafgaand aan modelkalibratie. Hierdoor worden waarnemingen in gronden met een relatief laag organisch-stofgehalte, d.w.z. minder dan 5%, die vaak voorkomen in gebieden met akkerland, zwaarder meegewogen in de kalibratieprocedure. De SoilGrids 2017 organische-stofkaart voor Europa werd ook vergeleken met de organische-stofkaart die gemaakt werd op basis van de EFSA dataset. Vergelijking met de LUCAS dataset liet zien dat SoilGrids 2017 het organische-stofgehalte over het algemeen enigszins overschatte, maar deze overschatting was veel kleiner dan die van de EFSA dataset.

Om de kaart voor het organische-stofgehalte in de bodem zoals gemaakt met het SoilGrids model verder te verbeteren werd een op maat gemaakte procedure ontwikkeld met een aantal aanpassingen in de kalibratieprocedure. Ten eerste werd het gebied waarvoor voorspellingen werden gedaan beperkt tot gebieden met akkerland als landgebruik, volgens de Corine landgebruikskaart van 2012. Daarnaast werd het SoilGrids model gekalibreerd met alleen waarnemingen van het organische-koolstofgehalte in akkerlanden. Aangezien het SoilGrids model het organisch-koolstofgehalte voorspelt, werd een conversiefactor gebruikt om het resultaat uit te drukken als een organisch-stofgehalte. De aldus met SoilGrids op maat gemaakte kaart met voorspellingen voor het organisch-stofgehalte in de bodem van akkerlanden in Europa werd aangeduid als SoilGrids 2018. Deze kaart werd gemaakt op basis van meer dan 40 000 waarnemingen van het organisch koolstofgehalte in de bodem en 200 kaarten van verklarende omgevingsvariabelen vanuit verschillende domeinen, zoals klimaat, ecologie, geologie, landgebruik, geomorfologie, vegetatie en water. Kaarten voor het organische-stofgehalte in de bodem werden gegenereerd met behulp van de machine-learning algorithmen Random Forest en Gradient Boosting, alsook met een Ensemble model dat een gewogen gemiddelde neemt van de resultaten van de Random Forest en Gradient Boosting modellen. Organische-stofkaarten werden gemaakt voor vier bodemdiepten, nl. 0.15 m, 0.45 m, 0.8 m en 1.2 m beneden maaiveld. De resultaten lieten zien dat het voorspelde gehalte sterk afneemt met de diepte, waarbij de sterkste afname optreedt tussen 0.15 en 0.45 m.

Voor de beoordeling van het risico op uitspoeling naar het grondwater is het van belang om het organische-stofgehalte in de range van enkele procenten (2-4%) zo nauwkeurig mogelijk te voorspellen. Daarom werden net als bij de ontwikkeling van de nieuwe organische stofkaart voor GEOPEARL de organische-stofgehalten eerst getransformeerd naar de log-schaal voorafgaand aan de verwerking van deze gegevens door SoilGrids. De nauwkeurigheid van de resulterende SoilGrids 2018 voorspellingen werd vervolgens bepaald met een tienvoudige kruisvalidatie. Hierbij worden de waarnemingen gesplitst in tien verschillende subsets van ongeveer gelijke grootte, waarbij steeds negen subsets worden gebruikt voor modelkalibratie en één subset voor validatie. De berekende Mean Error voor de RF en GB modellen bedroeg respectievelijk 0.58% en 0.43% organische stof. De Root Mean Squared Error waarden varieerden van 4.59% tot 5.08%. Voor de deelverzameling van data met een organisch-stofgehalte beneden 10% werd de berekende Mean Error veel kleiner, respectievelijk 0.09% en 0.01% voor het RF en GB model. De bijbehorende Root Mean Squared Error waarden werden berekend op respectievelijk 1.32% en 1.55%. Door kalibratie met gegevens op logschaal werd de nauwkeurigheid van het voorspelde organische-stofgehalte voor de deelverzameling van data met een organisch-stofgehalte beneden de 10% dus veel beter. Beperking van de data tot die met een organisch-stofgehalte < 5% in plaats van < 10% leverde maar een relatief kleine verdere verbetering op.

De voorspellingen met het Random Forest model waren nauwkeuriger dan die verkregen met het Gradient Boosting model. De resultaten van het Ensemble model waren iets beter dan die verkregen met het Random Forest model en dit model werd vervolgens gebruikt om de organische-stofkaart voor akkerlandgronden in Europa te maken. De resultaten van de kruisvalidatie wezen uit dat met SoilGrids 2018 en de gebruikte kalibratiestrategie voldoende nauwkeurige voorspellingen gemaakt worden van het organisch-stof gehalte in deze gronden.

Op basis van de resultaten van het Ensemble model gekalibreerd met organisch-stof gehalten op de logschaal werden kaarten voor de EU gegenereerd voor het organische-stofgehalten op 0.15 m, 0.45 m, 0.80 m en 1.2 m diepte. De kaart voor 0.15 m diepte werd vergeleken met de EFSA 2012 kaart. Over het geheel genomen was het met SoilGrids 2018 voorspelde organische-stofgehalte aanzienlijk lager dan dat volgens de EFSA 2012 kaart, met name voor centraal Europa. Dit verschil kan verklaard worden door het feit dat de EFSA 2012 kaart organische stofgehalten in gronden met alle typen landgebruik in rekening brengt, terwijl SoilGrids 2018 alleen die gebieden beschouwd met akkerland als landgebruik.

De met SoilGrids 2018 verkregen organische-stofkaarten voor de EU werden ook vergeleken met de SolilGrids 2017 kaarten. De SoilGrids 2018 voorspellingen waren over het algemeen lager dan de SoilGrids 2017 voorspellingen, dankzij modelkalibratie op log-getransformeerde organische-stofgehalten en gebruikmaking van alleen waarnemingen in akkerland.

Uit validatie met behulp van de LUCAS dataset bleek dat SoilGrids 2017 en met name EFSA 2012 het organisch-stofgehalte van akkerlanden in Europa systematisch overschatten, terwijl dat niet het geval is bij SoilGrids 2018.

De resultaten van deze studie laten zien dat met SoilGrids 2018 een verbeterde kaart voor het organische-stofgehalte in de bodem voor de EU is gemaakt in vergelijking met de EFSA 2012 en SoilGrids 2017 kaarten. De SoilGrids 2018 kaart verkregen in deze studie zou gebruikt kunnen worden als gegevensbron bij het maken van een bodemschematisatie voor een ruimtelijk-gedistribueerd model voor de beoordeling van de uitspoeling van pesticiden naar het grondwater op EU niveau. Een dergelijke schematisatie wordt momenteel uitgewerkt door de SETAC Werkgroep voor Spatially Distributed Leaching Modelling (SETAC-SDLM).

SoilGrids 2018 kan nog verder verbeterd worden door het toevoegen van nieuwe waarnemingen van het organisch-stofgehalte in de bodem voor de kalibratie procedure. Hierbij is het van belang te letten op de ruimtelijke verdeling van de waarnemingen. Deze verbetering kan vooral belangrijk zijn in gebieden met relatief weinig waarnemingen. Ook kan de set omgevingsvariabelen ten behoeve van de voorspellingen uitgebreid en verbeterd worden.

1 Introduction

1.1 Assessment of leaching to the groundwater

Leaching of crop protection products to groundwater is evaluated in procedures for the registration of these products, both at the national level and the European level. In the EU the leaching of crop protection products to groundwater is assessed according to EC regulation 1107/2009. For the assessment of the leaching potential in the first tier, the FOCUS Groundwater scenarios have been developed. These scenarios intend to represent vulnerable cases in different climatic zones of the EU. In tier 3b and tier 4 assessments, a spatially-distributed leaching model is to be used. The use of such a model requires high-resolution spatial data on soil properties of good quality. In the Netherlands, a new decision tree for the evaluation of the leaching potential of pesticides has been adopted in 2004. In this decision tree, the FOCUS Kremsmuenster scenario is used in tier 1. In tier 2, the leaching potential is assessed in the area of use. The target concentration for this assessment is defined as the median leaching concentration of a pesticide or its relevant metabolites at 1.0 m depth over a period of 20, 40 or 60 years for annual, biennial or triennial applications, respectively, within 90% of the area of use, which should not exceed the drinking water limit of 0.1 µg/L. To facilitate the calculation of this target value, the spatially-distributed model GeoPEARL has been developed, based on spatial data of soil type, climate, hydrotype (coupled to bottom boundary) and land use. Since the first release of GeoPEARL in 2004 (Tiktak et al., 2003; Van der Linden et al., 2004), several updates have been released in support of the Dutch registration procedure for plant protection products. The input of the GeoPEARL model consists of GIS data files representing a schematisation of the climate, soils, drainage systems, land use and crops for the agricultural area in the Netherlands. Assessments of the leaching to groundwater using GeoPEARL could also be done for other countries, zones or regions. This would require a preparation of a set of schematisation files representing the GeoPEARL inputs for the area of interest.

Soil organic matter (SOM) content is a key factor in the assessment of the leaching of plant protection products. Therefore it is important to obtain reliable data on the organic matter content in arable soils in the area of use of the crop protection product.

1.2 A new SOM map for Europe

To date, the EFSA dataset (EFSA, 2012) is being used as a source for data on the SOM content at European scale (e.g. EFSA, 2017). An important disadvantage of this dataset is that it appears to be biased for arable soils, possibly because it does not explicitly account for land use differences in the Corine Land Cover map as included in the EFSA dataset. As the data included in creating this dataset also relate to grassland and forest soils, which typically have much higher SOM content than arable soils, the EFSA dataset is likely to overestimate the SOM content in arable soils. Therefore, there is a need for an improved dataset for the SOM content at the European scale.

Over the past years ISRIC has developed the SoilGrids model to map soil organic carbon (SOC) content based on soil organic carbon measurements and correlated environmental variables (Hengl et al., 2017). To explore whether the SoilGrids methodology could be useful to obtain an improved SOM map for Europe, the existing SoilGrids SOM maps for the Netherlands was compared with the 2017 Dutch SOM map for GeoPEARL (Van den Berg et al., 2017). The results are shown in Figure 1.1. The SOM content was obtained by multiplying the SoilGrids soil organic carbon content by a factor 2.0, which has been found to be applicable for the conversion of SOC content to SOM content in Dutch soils (Sleutel et al., 2007). In the southern half of the Netherlands, a large area of the SoilGrids SOM map indicates an organic matter content between 3 and 5% in the topsoil. In the northern half of the country organic matter contents are overall even higher, mostly in the range of 5 to 10%. When comparing the map derived from SoilGrids with the 2017 Dutch soil organic matter map for GeoPEARL, the former is roughly more than 2.5% higher than the latter (see bottom left map in Figure 1.1).

A number of factors were identified that could contribute to the differences in the two SOM maps. SoilGrids is a global model which uses SOM measurements from across the globe, the vast majority of which is outside the Netherlands, to train the machine learning algorithm underlying SoilGrids. Therefore, SOC measurements from areas with similar environmental conditions as the Netherlands will have been used to derive SOM estimates in the Netherlands. The model that was developed to predict the SOM content for the 2017 Dutch SOM map for GeoPEARL takes a very different approach. It is based on a trend model using nine different soil types and an interpolation step based on 770 000 soil observations from within the Netherlands. Before calibration of the model, the SOM contents were transformed to contents on a log scale. This approach was adopted because it allows a more accurate prediction of SOM contents in the range typical for arable land. So these two models use different SOM measurements, different explanatory variables and different (geo-)statistical models to predict SOM content.

One important factor that also may have contributed to the systematic difference between the SoilGrids SOM map and the 2017 Dutch SOM map for GeoPEARL is that SoilGrids does not focus explicitly on arable land. As mentioned above, organic matter content in the topsoil of arable land is typically much lower than that in grassland soils. Land use is considered in SoilGrids, using a global land cover map for 2010 with a resolution of 30 m (Chen et al., 2015). But even though SoilGrids distinguishes land use, it uses soil organic carbon observations from all land use types for model calibration and soil organic matter prediction. In contrast, the 2017 Dutch SOM map for GeoPEARL presented in Figure 1.1 (top left) refers to the SOM content assuming arable land use. This is because pesticides are mostly only applied to arable land and hence GeoPEARL needs SOM content estimates for arable land. For this comparison a multiplication factor of 2.0 was used for the transformation of SOC predicted by SoilGrids to SOM.

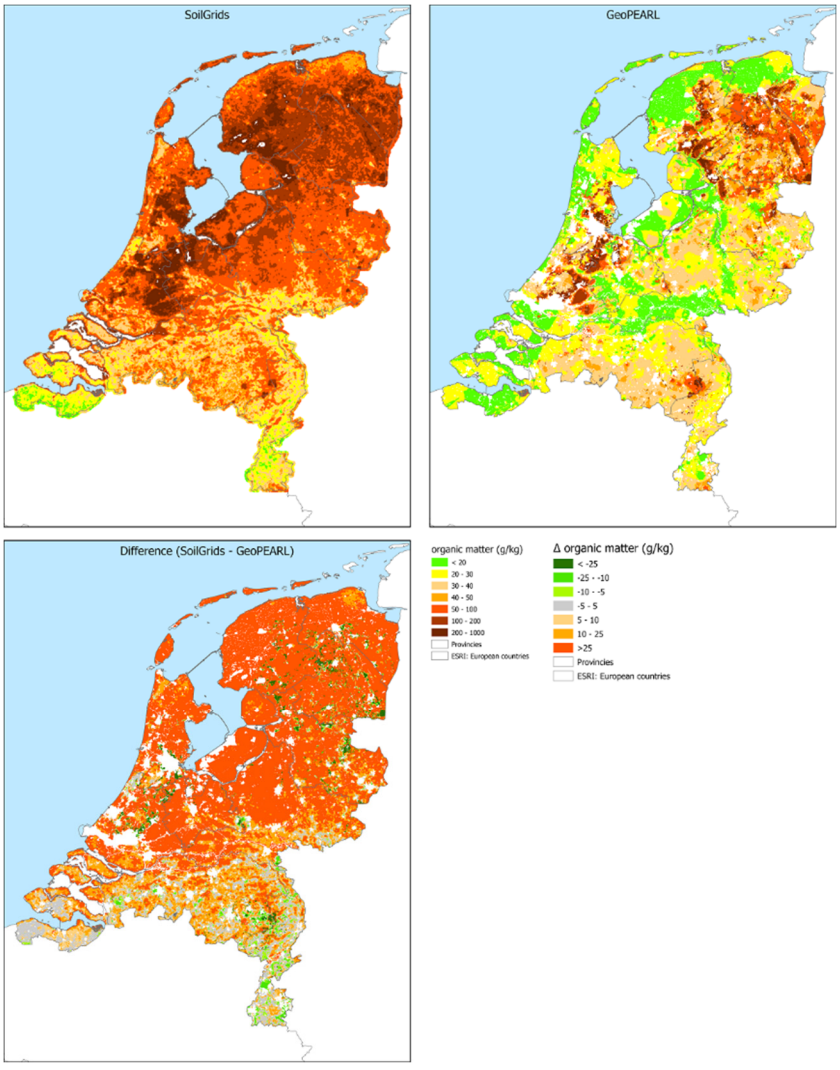

Figure 1.1 *Soil organic matter map at a depth of 0.15 m using SoilGrids (top left), GeoPEARL_SOM 2017 (top right) and for the differences (bottom left). Soil organic matter expressed as g/kg.*

Based on the considerations listed above a better comparison between SoilGrids and the 2017 Dutch SOM map for GeoPEARL could be made by limiting the comparison to areas in the Netherlands with arable soils. In addition, SoilGrids could be trained with SOM measurements from arable land only.

Also, since SoilGrids uses input data on soil organic matter from the Netherlands that were converted to soil organic carbon by division by 1.724 instead of 2.0, it seems more appropriate to transform the soil organic carbon map to a soil organic matter map by multiplication by 1.724 instead of 2.0. All the possible modifications to the SoilGrids modelling approach will be addressed in this report.

The potential of SoilGrids to provide a SOM map for Europe was further explored by comparison of the existing SoilGrids map with the EFSA map. The SOM map for the European Union derived from SoilGrids, using conversion factor 2.0 as before, is presented in Figure 1.2 (top left). The soil organic matter map using the EFSA data is also shown in this figure (top right). Overall, the soil organic matter content derived from SoilGrids for Central and Northern Europe is generally 1% to 2.5% lower than the estimate from EFSA data (in some places the difference is even larger). One of the shortcomings of the EFSA dataset is that it does not use land use as an explanatory variable, while SoilGrids does. As a result the map generated with SoilGrids may be expected to be closer to reality than the EFSA map. But note that the SoilGrids 2017 SOM map shown in Figure 1.2 was obtained using a model that was calibrated on soil organic carbon observations from all land uses. Also at the European scale, a more accurate SOM map for arable land may likely be obtained by calibrating the SoilGrids model using soil data from arable land only.

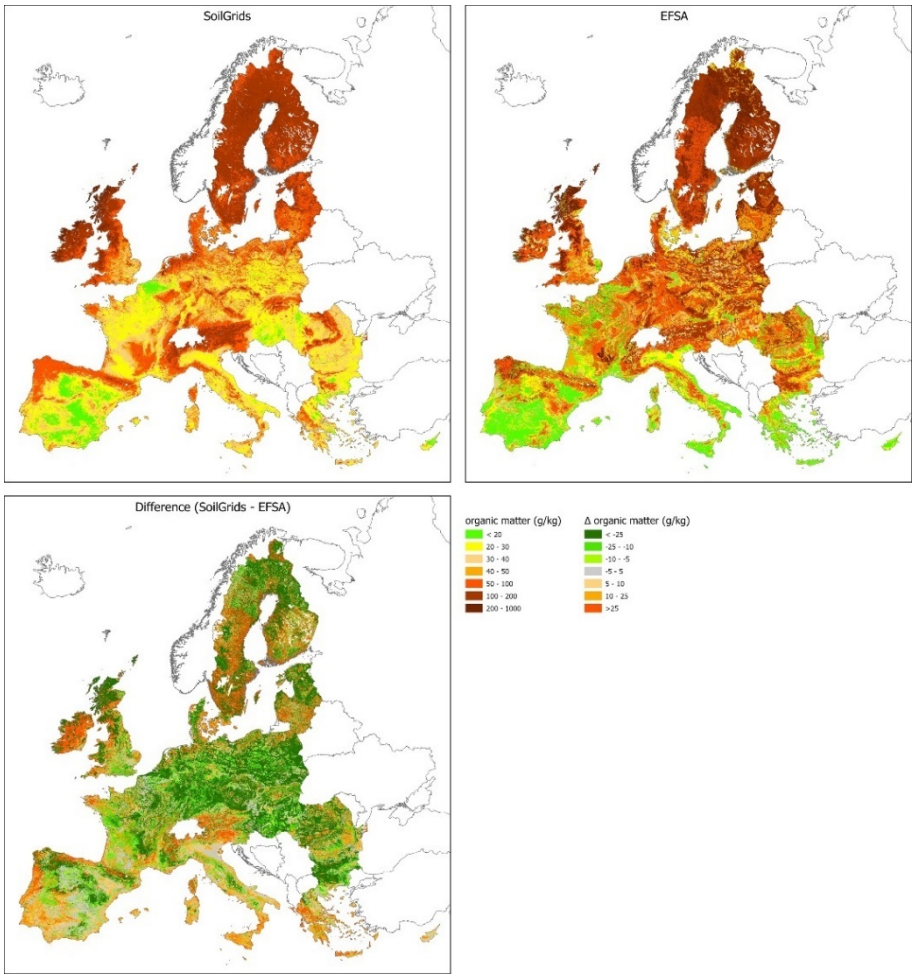

Figure 1.2 *EU soil organic matter map for the top 0.3 m using EFSA data (top left), SoilGrids (top right) and a map showing the differences (bottom left). Soil organic matter expressed as g/kg.*

1.3 Goals and outline of this work

The aim of this study is to obtain an improved soil organic matter map that could be used to derive a soil schematisation for a spatially-distributed model to assess the leaching of pesticides to groundwater at the EU level. In light of the reconnaissance reported above, it seems that an improved soil organic matter map for the purpose of the assessment of leaching to groundwater at the EU level could be obtained using the SoilGrids model by introducing a filter on input data, so that only soil data from sites on arable land in the EU are considered for SoilGrids calibration, thus restricting the area of interest to the area being classified as arable land. An overlay with the CORINE land use map could be used to select only data points that are in arable land, and to limit comparison between SoilGrids and EFSA maps to arable land. However, before embarking on adapting the SoilGrids methodology to the needs of the assessment of leaching to groundwater in the EU it is useful to compare the existing SoilGrids 2017 maps with the 2017 Dutch soil organic matter map for GeoPEARL and the EFSA map in more detail. Therefore, the present study aims first to check whether the SoilGrids methodology is useful by making these extended comparisons, before applying the SoilGrids method at the EU scale and assess its performance.

The results of this extended comparison are presented in Chapter 2. In this chapter, a value of 1.724 instead of 2.0 is used to transform SoilGrids soil organic carbon to soil organic matter, which is the reciprocal of the "Van Bemmelen" factor of 0.58. In addition maps are clipped to arable land in the Netherlands and EU. Further, summary statistics of the differences are computed and validation statistics calculated.

In Chapter 3 the procedure to obtain a tailored SoilGrids map for the EU is presented, by calibrating SoilGrids with soil organic matter observations on arable land in the EU (henceforth referred as SoilGrids 2018). This procedure is applied to obtain maps for the soil organic matter content at different depths for the arable lands of the EU. The accuracy and quality of the soil organic matter maps produced in this tailored approach is assessed in Chapter 4 by computing validation statistics as well as by performing a cross-validation strategy. In Chapter 5 the new SoilGrids 2018 SOM map is compared with existing datasets such as the EFSA dataset. Finally, conclusions and recommendations for the use of SoilGrids to produce SOM maps for arable lands in Europe are presented in Chapter 6.

2 Assessing feasibility of using SoilGrids for NL and EU soil organic matter maps

In this chapter the SOM maps produced with SoilGrids 2017 (Hengl et al., 2017) for the Netherlands are compared with the 2017 Dutch soil organic matter map for GeoPEARL (van den Berg et al., 2017) and that for the EU with the EFSA SOM map for the EU (EFSA, 2012). This exercise informs on general trends and systematic differences between these maps to assess the potential for the use of the SoilGrids soil organic matter map for Europe in the groundwater leaching assessment at the EU level (EC 2014).

For brevity of text, the three maps are henceforth referenced in short as GeoPEARL_SOM 2017, SoilGrids 2017 and EFSA SOM.

2.1 SoilGrids 2017 versus GeoPEARL_SOM 2017

2.1.1 Direct map comparison

Firstly, the SoilGrids 2017 predictions were compared with the GeoPEARL_SOM 2017 map for a depth of 0.15 m. As in SoilGrids only organic carbon content is predicted, a SOM map was obtained with a multiplication factor of 1.724. The comparison was limited to the area in NL with soils under arable land as specified in the land use map included in GeoPEARL. This land use map is based on LGN 4 (published in 2001).

Figure 2.1 shows the histograms of both maps for arable land in the Netherlands. The predictions for the soil organic matter content are generally higher with SoilGrids 2017 than with GeoPEARL_SOM 2017. Whereas most values on the latter map are below 50 g/kg, the SoilGrids 2017 predictions are frequently above that level. Further, the SoilGrids 2017 map has a larger number of predictions above 100 g/kg. This can also be seen in the summary statistics presented in Table 2.1.

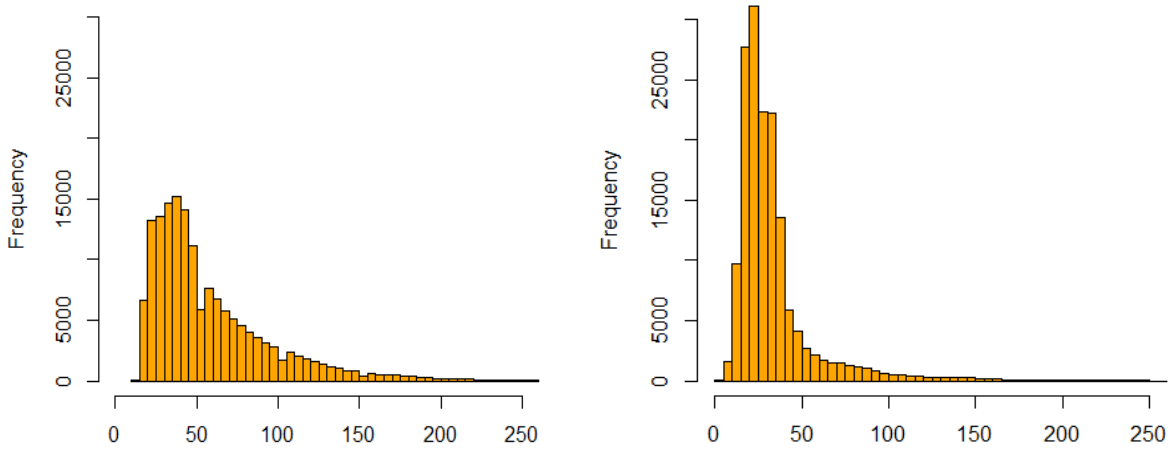

Figure 2.1 *Histograms of soil organic matter contents (g/kg) for the top at a depth of 0.15 m for arable land in the Netherlands for SoilGrids 2017 (left) and GeoPEARL_SOM 2017 (right).*

Table 2.1 *Summary statistics of organic matter content (g/kg) at a depth of 0.15 m on arable land for SoilGrids 2017 and GeoPEARL_SOM 2017, and their difference.*

	SoilGrids 2017	GeoPEARL_SOM 2017	Difference (SoilGrids 2017 – GeoPEARL_SOM 2017)
Minimum	13.8	0.7	463.0
1st Quantile	32.8	20.1	4.3
Median	46.6	26.9	19.5
Mean	58.9	34.2	24.7
3rd Quantile	74.1	36.1	39.0
Max	460.3	655.3	363.9
Standard deviation	38.8	30.2	34.0
Skewness	1.86	6.62	0.20

Figure 2.2 shows the two soil organic matter maps. Both maps show comparatively high soil organic matter contents in the North-East (Groningen and Drenthe Provinces) and the Flevoland polders in the Centre and lower soil organic matter contents in the South-West (Zeeland Province). Figure 2.3 presents the spatial distribution of the differences between the two maps (SoilGrids 2017 – GeoPEARL_SOM 2017). This figure shows that SoilGrids 2017 indicates substantial higher SOM values than GeoPEARL_SOM 2017 in the provinces of Flevoland and Zeeland, in the east of Groningen and close to the Wadden Sea, but in general the two maps differ across the entire country. Most differences are positive, reflecting higher predictions in SoilGrids. There are a few small patches where the GeoPEARL_SOM 2017 map has much higher values than SoilGrids, e.g. the Southern part of Limburg Province in the South-East of the Netherlands.

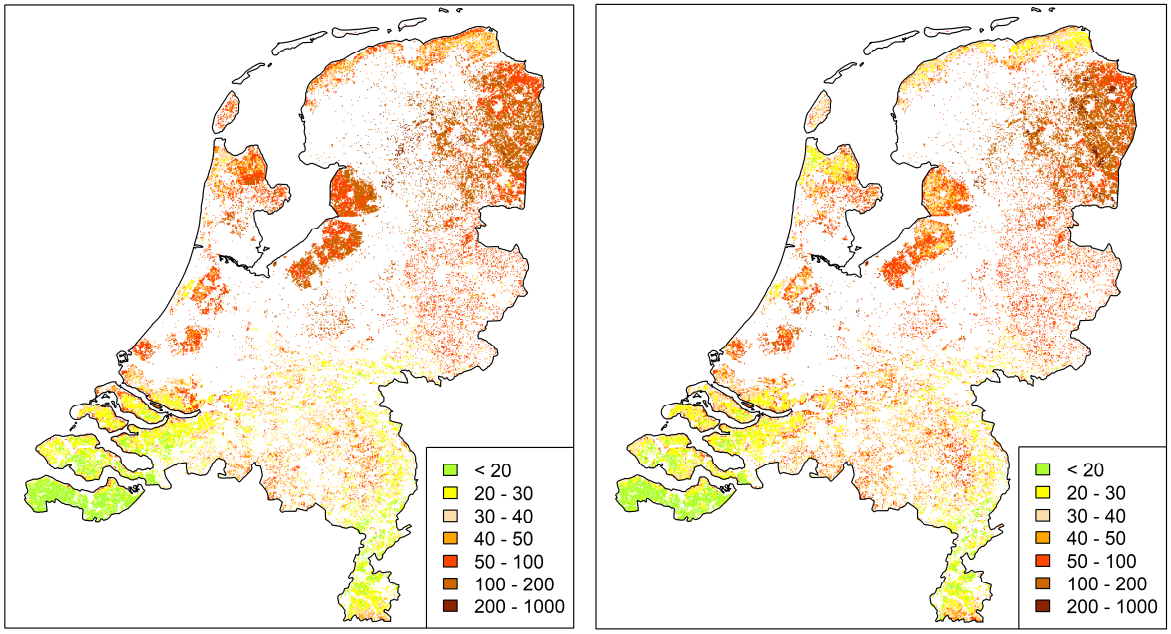

Figure 2.2 *The soil organic matter map (g/kg) for 0.15 m depth using SoilGrids 2017 (left) and GeoPEARL_SOM 2017 (right).*

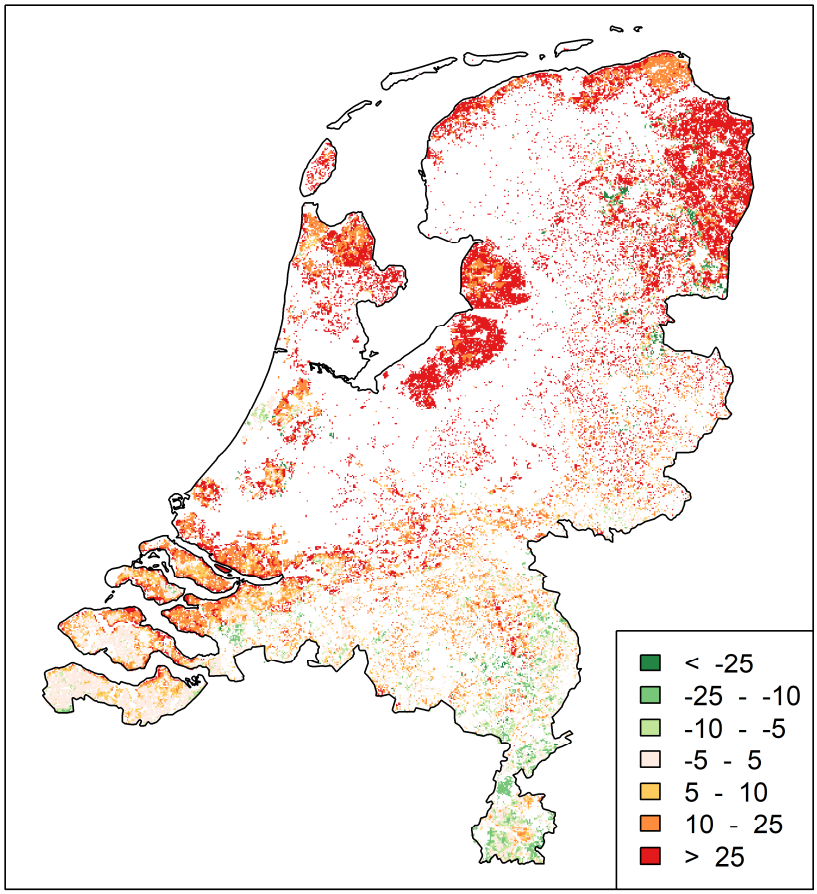

Figure 2.3 *Spatial distribution of the differences (g/kg) between SoilGrids 2017 and GeoPEARL_SOM 2017 on arable land for 0.15 m depth. Difference positive if SoilGrids 2017 > GeoPEARL_SOM 2017.*

The scatter density plot in Figure 2.4 provides a further perspective into the differences between the two maps. In cases where SoilGrids has values above 300 g/kg, the values predicted with the GeoPEARL_SOM 2017 map are consistently lower. Likewise, the GeoPEARL_SOM 2017 predictions greater than 400 g/kg disagree much with SoilGrids 2017, which are mostly below 300 g/kg. The majority of observations are at values smaller than 50 g/kg, and here the SoilGrids over-prediction is apparent since the cluster of high densities is clearly centred above the 1:1 line. The histogram of the differences as shown in Figure 2.4 (right) shows a positive mean and median as well as a positive skewness.

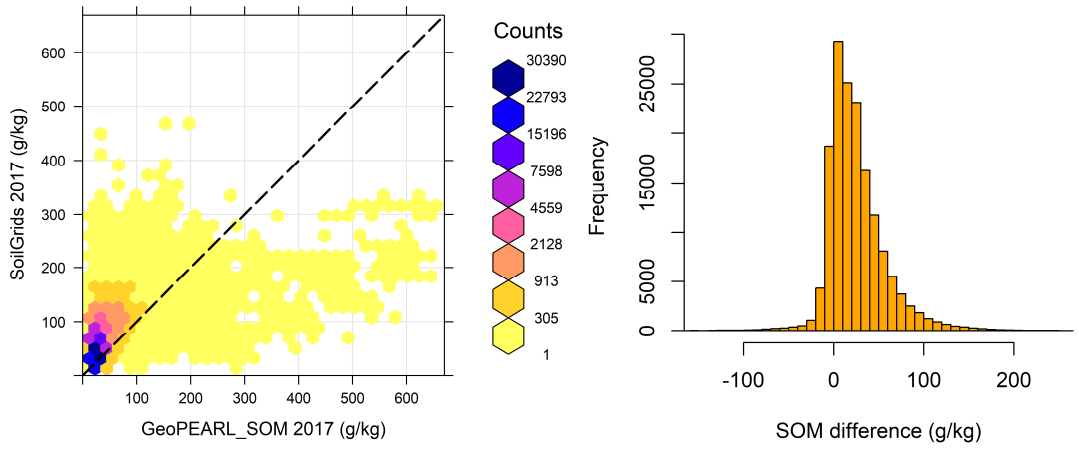

Figure 2.4 *Bivariate density plot of soil organic matter predictions (left) and histogram of differences between the two soil organic matter maps (right). Difference positive if SoilGrids 2017 > GeoPEARL_SOM 2017.*

2.1.2 Map validation using LSK and SoilGrids input data

For the statistical validation LSK (Landelijke Steekproef Kaarteenheden) data were used (Finke et al., 2001; Visschers et al, 2007). An arable land mask was applied to limit the LSK data to arable land areas as specified in the schematisation of land use in GeoPEARL (see Section 2.1.1). The application of this filter reduces the number of validation data to 392. The validation metrics used were:

- Root mean squared error (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}$
- Mean absolute error (MAE): $\frac{1}{n}\sum_{i=1}^{n}|O_i - P_i|$
- Mean error (ME): $\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)$
- Model Efficiency Coefficient (MEC, in %): $\left(1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}\right) \times 100\%$

where n is the size of the validation set, O_i the i-th observation, P_i the i-th prediction, and $\overline{O} = \frac{1}{n}\sum_{i=1}^{n} O_i$ the means of the observations.

The validation statistics for the prediction of soil organic matter using SoilGrids 2017 and the GeoPEARL_SOM 2017 map as based on the LSK observations on arable land (n = 392) are shown in Table 2.2, while scatter density plots are shown in Figure 2.5. To simplify the analysis we did not account for the fact that the LSK data are a stratified simple random sample and treated all observations equally when computing the validation metrics. As it cannot be assumed that the LSK stratified sample is self-weighing, improved validation statistics could have been obtained by taking the weights into account. The results show that SoilGrids 2017 somewhat overestimates the soil organic matter content at the LSK sites, whereas GeoPEARL_SOM 2017 somewhat underestimates it. The Model Efficiency Coefficient (MEC, Nash and Sutcliffe, 1970) indicates whether the model is perfect (MEC equal to 100%) or as good as taking the average of the observations (MEC equal to 0%). Both maps have a poor Model Efficiency Coefficient, about 6% for the GeoPEARL_SOM 2017 map to about -11% for the SoilGrids 2017 map. The Root Mean Squared Error for both maps is similar.

It should be noted that the validation statistics were obtained using the actual soil organic matter data. If a log-transformation was done on the observation data prior to computing validation statistics then these statistics might improve. Without transformation, observations in soils with a high organic matter content affect the outcome of the validation statistics more. In fact, the GeoPEARL_SOM 2017 maps was obtained after calibration of the log-transformed soil organic matter (van den Berg et al., 2017). Note that the validation procedure for the GeoPEARL_SOM_2017 map was a cross-validation as described by Goovaerts (1997).

Table 2.2 *Summary statistics of the comparison of SoilGrids soil organic matter predictions and those for the new GeoPEARL soil organic matter map with LSK data (n=392).*

	Mean Error (g/kg)	Root Mean Squared Error (g/kg)	Model Efficiency Coefficient (%)
SoilGrids	-6	40	-10.9
SOM GeoPEARL	18	37	6.4

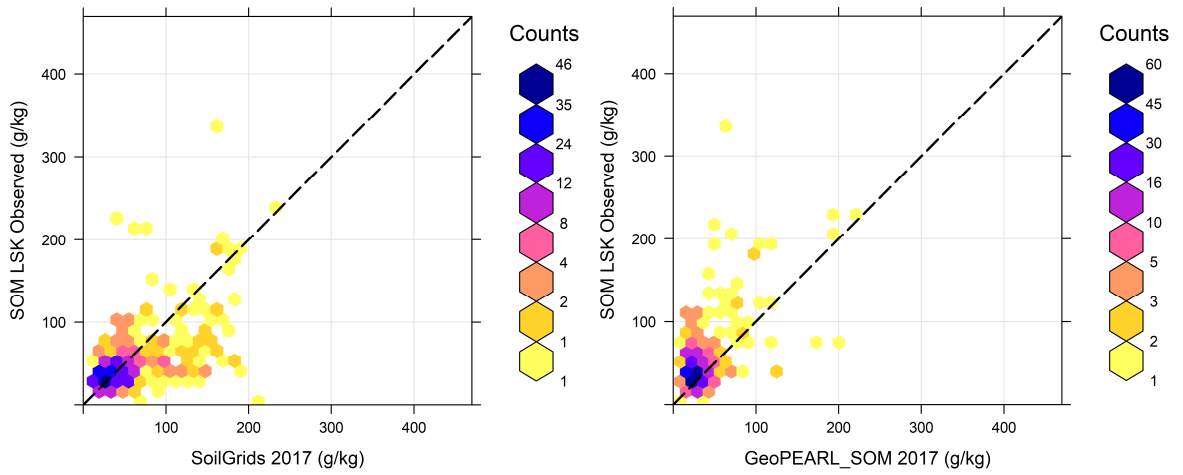

Figure 2.5 *Scatter density plots for comparison of LSK data (n=392) with the SoilGrids 2017 soil organic matter map (left) and the GeoPEARL_SOM 2017 map (right).*

The model predictions for the organic matter content using SoilGrids 2017 and those for the GeoPEARL_SOM 2017 map were also compared to the SoilGrids input data on arable land in the Netherlands (n=268). These data are not part of LSK, but NL data collected over the years and added to the ISRIC database. Note that the SoilGrids 2017 maps are not independent from the SoilGrids input data and that this may bias the validation results. Table 2.3 and Figure 2.6 show the results of this comparison.

Again, SoilGrids 2017 overestimates somewhat the soil organic matter content measured at the observation sites, whereas the GeoPEARL_SOM 2017 map underestimates it. This bias in the GeoPEARL_SOM 2017 map is caused by the log-transformation. Van den Berg et al. (2017) have shown that the mean error of log-transformed organic matter contents is generally close to zero, so the predictions at the log-scale are unbiased. However, after conventional back-transformation using the antilog, a bias is introduced. Using these data for validation the Model Efficiency Coefficient is substantially higher than that obtained using the LSK data. This is confirmed by the reduction of scatter between the scatter density plots shown in Figure 2.6 compared to the scatter density plots shown in Figure 2.5. A possible explanation for the improved Model Efficiency when using the SoilGrids NL data might be that these data were used to calibrate the SoilGrids machine learning model, although this cannot explain why SOM GeoPEARL also performs better. Interestingly, the RMSE values presented in Table 2.3 are substantially larger than those in Table 2.2. This suggests that model performance is worse when evaluated on the SoilGrids NL data. The explanation for these seemingly contradictory results is that the SoilGrids NL data have a much larger variance than the LSK data, mainly due to one extreme value of 951 g/kg (not shown in Figure 2.6).

Table 2.3 *Summary statistics of the comparison of SoilGrids 2017 and the GeoPEARL_SOM 2017 map with SoilGrids NL data (n=268). Mean error defined as mean of observations minus mean of predictions, so that a negative value indicates overestimation.*

	Mean Error (g/kg)	Root Mean Squared Error (g/kg)	Model Efficiency Coefficient (%)
SoilGrids	-6	52	47.9
SOM GeoPEARL	16	59	40.0

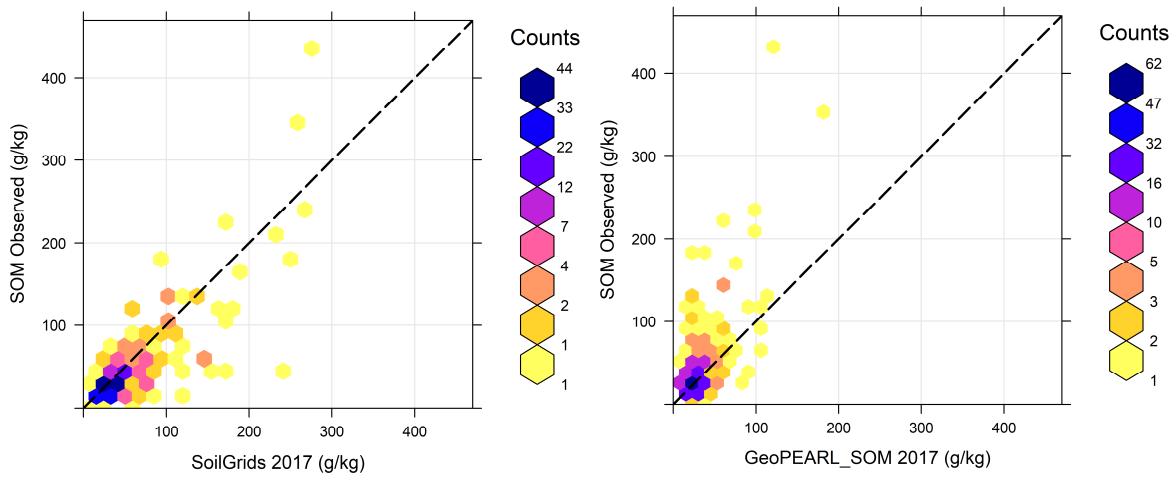

Figure 2.6 *Scatter density plots of NL SoilGrids observations (n=268) with the SoilGrids 2017 soil organic matter map (left) and the GeoPEARL_SOM 2017 map (right).*

2.2 SoilGrids 2017 versus EFSA SOM

A second comparison was conducted within the arable lands of the EU between the SoilGrids 2017 and the EFSA SOM maps. A multiplication factor 1.724 was again used to derive SOM for the SOC prediction in SoilGrids 2017. In addition, the average soil organic matter content for the top 0.3 m was computed with the SoilGrids 2017 map to match the top soil interval depth in the EFSA SOM map. Both maps are presented in Figure 2.7. SOM predictions using SoilGrids 2017 are on average lower than those in the EFSA SOM map, in particular for Northern and Central Europe.

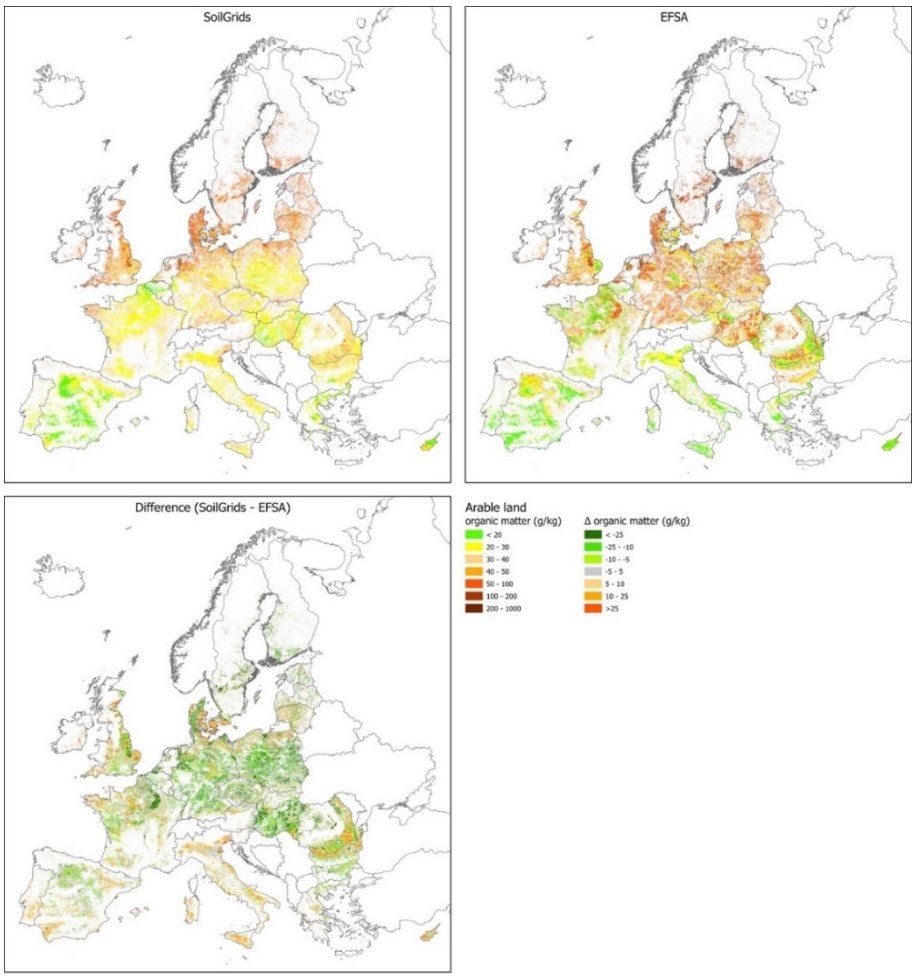

Figure 2.7 *Soil organic matter (g/kg) in the top soil of arable lands in the EU using SoilGrids 2017 (left) and the EFSA SOM map (right). Difference map also shown.*

The two maps were then compared with the LUCAS dataset (Orgiazzi et al., 2018). The LUCAS dataset includes measurements obtained from 45 000 samples of the top soil (0 to 0.2 m deep) collected across the continent. Only observations on arable land were used (clc12 codes 211 Non-irrigated arable land, 212 Permanently irrigated land and 213 Rice fields). Scatter density plots for this comparison are shown in Figure 2.8. It should be noted that part of the LUCAS dataset was used for calibration of the SoilGrids 2017 SOC map, so the validation test was not truly independent. However, the aim of this comparison was just a preliminary step towards the generation of a new SOM map, so the validation step was not very critical at this stage.

The results presented in Table 2.4 show that the EFSA SOM map overestimates the actual soil organic matter content at the LUCAS data points on arable land. On average the SoilGrids 2017 predictions agree well with the SOM contents measured at the LUCAS data points. The root mean squared error of the SoilGrids predictions is only half of that of the EFSA SOM map. Furthermore, the model efficiency of SoilGrids 2017 is much better than that of EFSA SOM.

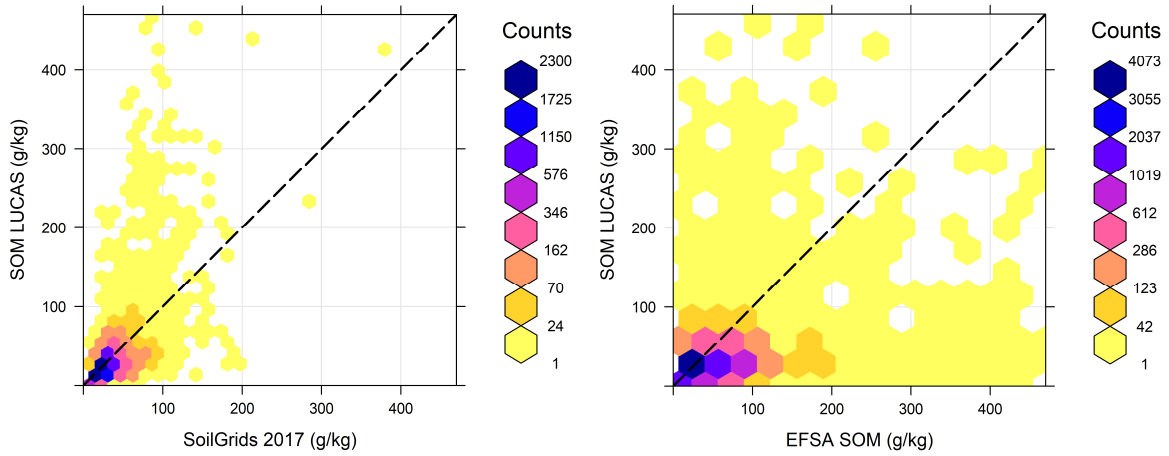

Figure 2.8 *Scatter density plots of LUCAS data on arable land (n=11 808) with the SoilGrids 2017 soil organic matter map (left) and the EFSA SOM map (right).*

Table 2.4 *Summary statistics of comparison of predicted soil organic matter contents using SoilGrids 2017 and the EFSA SOM map with LUCAS data on arable land (n=11 808). Mean error defined as mean of observations minus mean of predictions, so that a negative value indicates overestimation.*

	Mean Error (g/kg)	Root Mean Squared Error (g/kg)	Model Efficiency Coefficient (%)
SoilGrids	0	41	37.6
EFSA_OM_TOP	-25	81	-135.1

2.3 Discussion

For the Netherlands, based on the LSK data on arable land, the accuracy of SoilGrids 2017 is comparable to the GeoPEARL_SOM 2017 map. Both maps have a similar root mean squared error. The SoilGrids predictions have a negative mean error, so a systematic overestimation, whereas the GeoPEARL_SOM 2017 map has a positive mean error. Both maps have a poor model efficiency. It should be noted (see Section 2.1) that the GeoPEARL_SOM 2017 map has been obtained by calibration of the soil organic matter contents on a logscale (van den Berg et al., 2017). This was done to improve the accuracy of the model predictions in soils with low organic matter contents. The risk of leaching to groundwater increases with decreasing organic matter content and soils with arable land use have comparatively low organic matter contents.

When compared with the LUCAS dataset the SoilGrids 2017 map for the EU has no bias, which is not surprising considering that it used part of that dataset for model calibration (but note that validation statistics were only computed for arable land). In turn the EFSA map overestimates SOM content systematically, as shown by the negative mean error. Both the root mean squared error and the model efficiency show that the SoilGrids 2017 map is closer to the LUCAS data.

In view of the findings listed above, it is recommended to explore the development and application of a tailored SoilGrids soil organic matter map as input for a model to assess leaching at the European scale. In addition, a comparison needs to be made between the 'tailored' SoilGrids soil organic matter map with the 'default' SoilGrids soil organic matter map, in order to evaluate whether the 'tailored' soil organic matter map indeed has greater accuracy.

3 Development of a EU soil organic matter map for arable land with SoilGrids

ISRIC has been predicting Soil Organic Carbon (SOC) of the world's soils for several years with SoilGrids (Hengl et al., 2014; Hengl et al., 2017; Poggio et al., 2021). The name SoilGrids is used interchangeably referring to the computer model that performs the predictions and its map outputs. SoilGrids relies on modern machine learning methods, trained on existing soil profile sample data, largely drawing on the WoSIS dataset (Batjes et al., 2017). At the time this study was conducted WoSIS included the observations available in the LUCAS dataset in 2016, comprising topsoil properties gathered for some 20 000 locations across the continent.

In this project the SoilGrids code base is used to create a SOM map for the arable lands of Europe. The SoilGrids SOC model is trained specifically for this study area, with a quality assessment based on a cross-validation procedure (conducted prior to prediction). The final SOM map is obtained from the SOC prediction applying a multiplication by 1.724. The SoilGrids model was configured to make predictions at four different depths: 15 cm, 45 cm, 80 cm and 120 cm.

3.1 Study area

The study area is defined as the spatial union of the classes in the Corine Land-cover map of 2012 (Büttner, 2014) identified as arable land. The following Corine classes were selected: *Annual crops*, *Agriculture*, *Permanent crops* and *Rice*. Figure 3.1 presents the resulting study area.

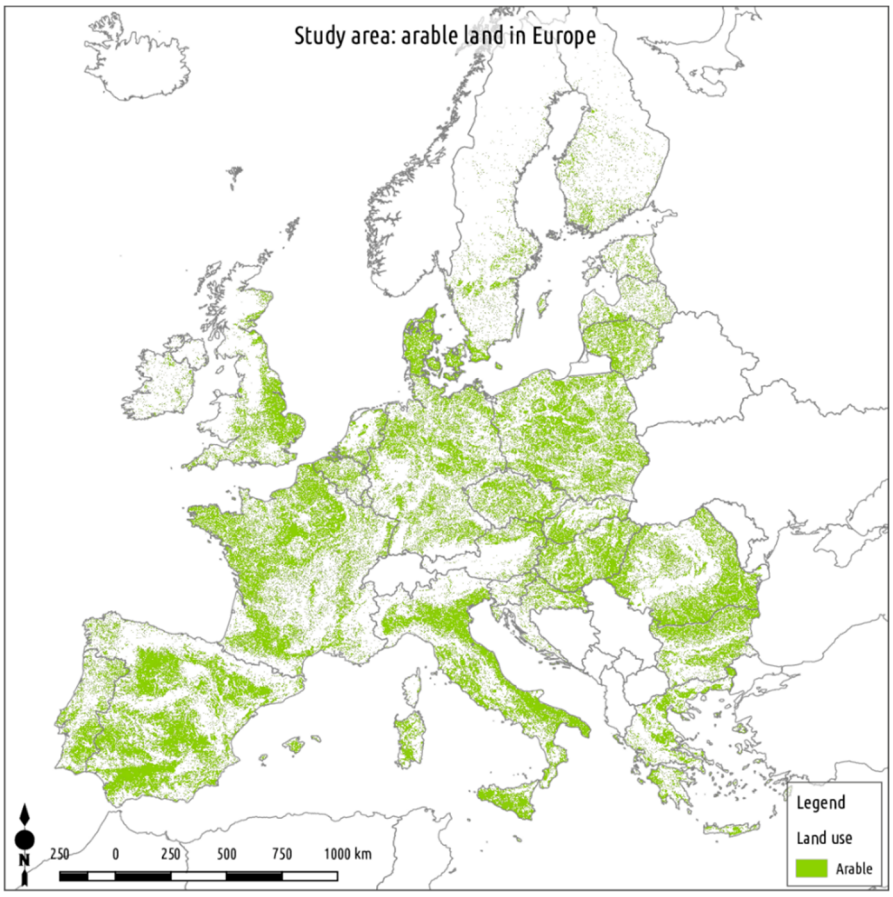

Figure 3.1 *Cartogram of the study area.*

3.2 SoilGrids tile system for arable land in Europe

SoilGrids performs predictions on a tiled structure of the study area. Each tile is an irregular polygon with a side of 54 minutes of arc. These tiles facilitate the parallelisation of predictions, allowing execution in a multi-CPU system. A tile set of similar characteristics was created for this study, encompassing the arable land areas shown in Figure 3.2. Adhering to the same 54 minutes of arc topology this tiling scheme could be used seamlessly with the 2017 SoilGrids code base. Figure 3.2 presents the tile structure created.

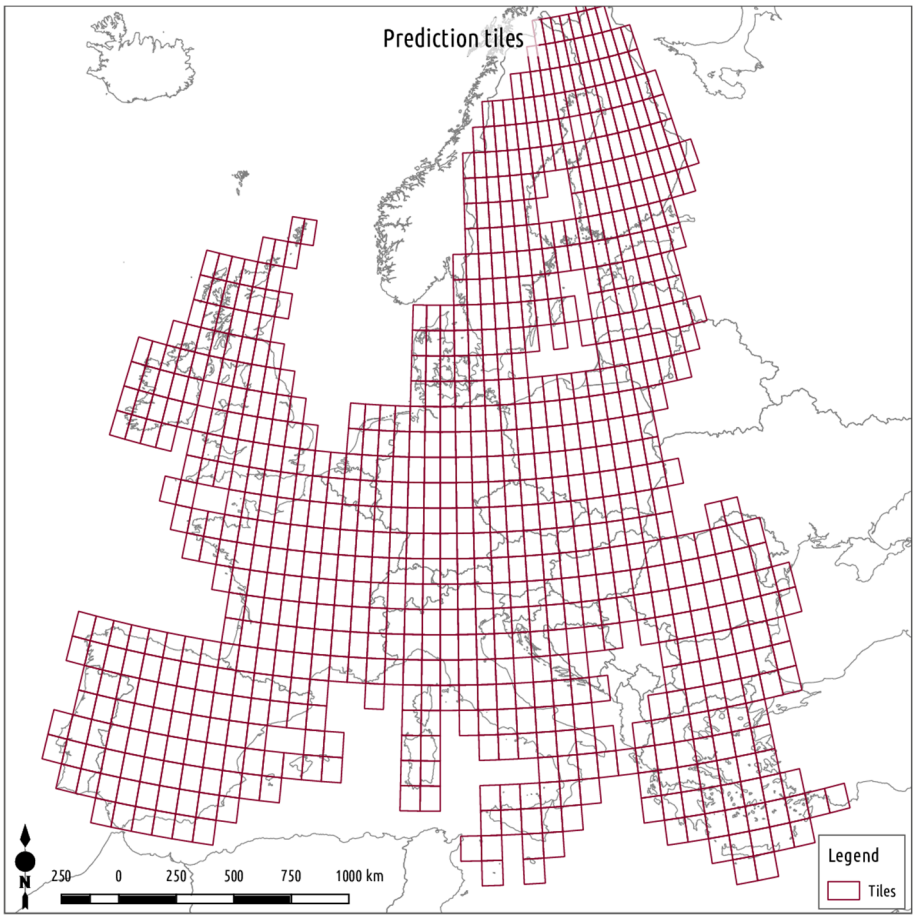

Figure 3.2 *Cartogram of the prediction tiles used in this study.*

3.3 Soil profile data used in this study

ISRIC's soil profile database was overlaid on the arable land layer obtained previously to select those profiles within the study area located on arable land. Figure 3.3 shows the locations of the 19 049 soil profiles selected for this study. There are pockets of higher spatial density, but most member states are well covered. Only three countries stand out as poorly covered: Ireland, Bulgaria and Romania. However, the climatic, morphological and geological characteristics of these regions are not dramatically different from other regions of Europe to prevent prediction. Since spatial location is not used as a prediction covariate, the predictive correlations found by the model in a particular region apply to other regions with similar environments.

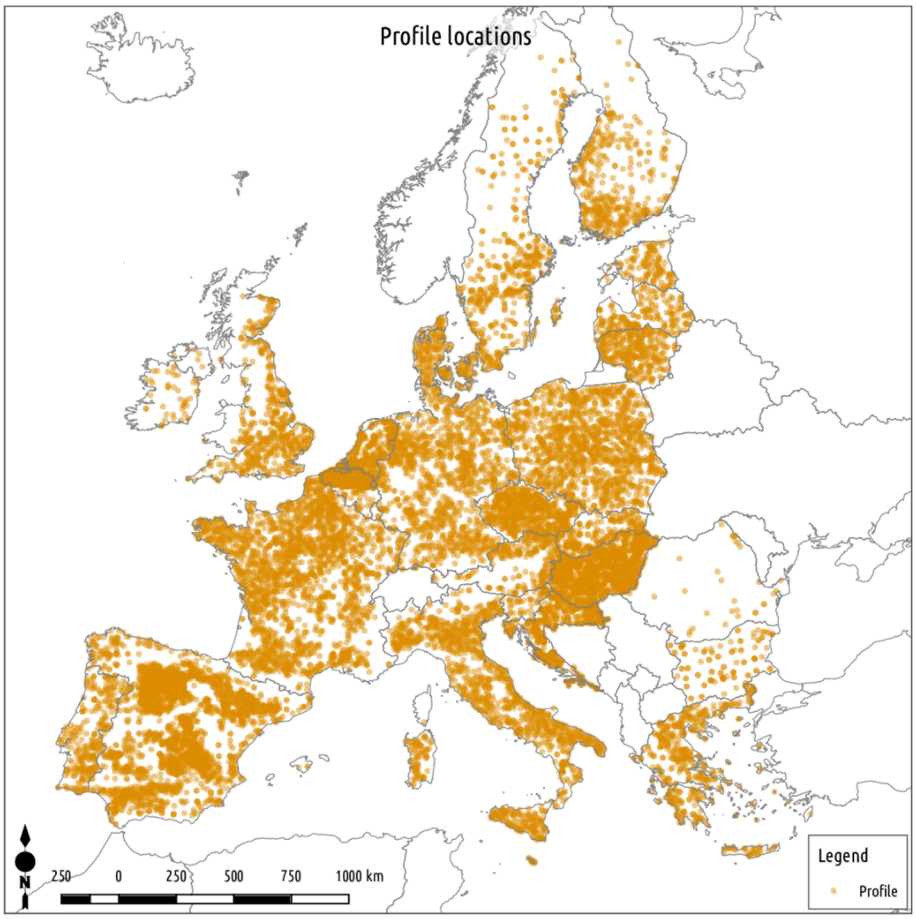

Figure 3.3 *Locations of soil profiles within the study area.*

The profiles selected within the study area were translated into 43 593 individual SOC observations, Figure 3.4 presents a histogram. Observations are highly concentrated in the range below 20 g/kg, in a very skewed distribution. However, there are various observations in the hundreds of g/kg. Table 3.1 presents summary statistics of the SOC observations.

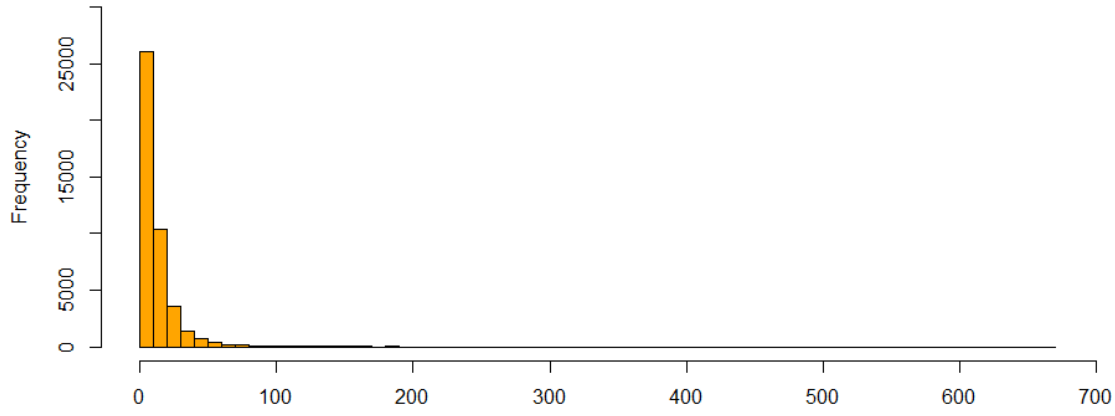

Figure 3.4 *Histogram of SOC observations (g/kg) used for model calibration.*

Table 3.1 *Summary statistics of SOC observations (g/kg).*

Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
0.00	2.15	7.300	13.78	15.08	665.60

In the original dataset these observations are assigned to a depth interval. For this study their position within the profile was reduced to the mid-point of the depth interval in order to make their use possible with the SoilGrids code base. Figure 3.5 presents a scatter density plot of SOC observations against the mid-point of the depth interval. As expected, SOC values at depths greater than 1 m are typically below 1% (or 10 g/kg).

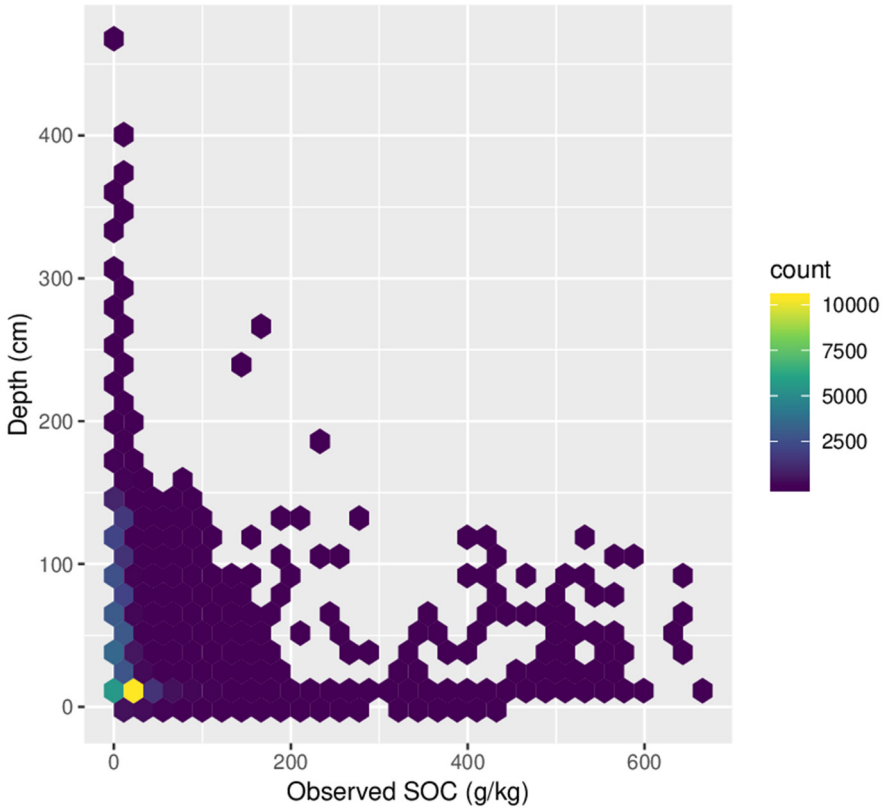

Figure 3.5 *Scatter density plot of SOC observations against depth of the SoilGrids training data on arable land in EU.*

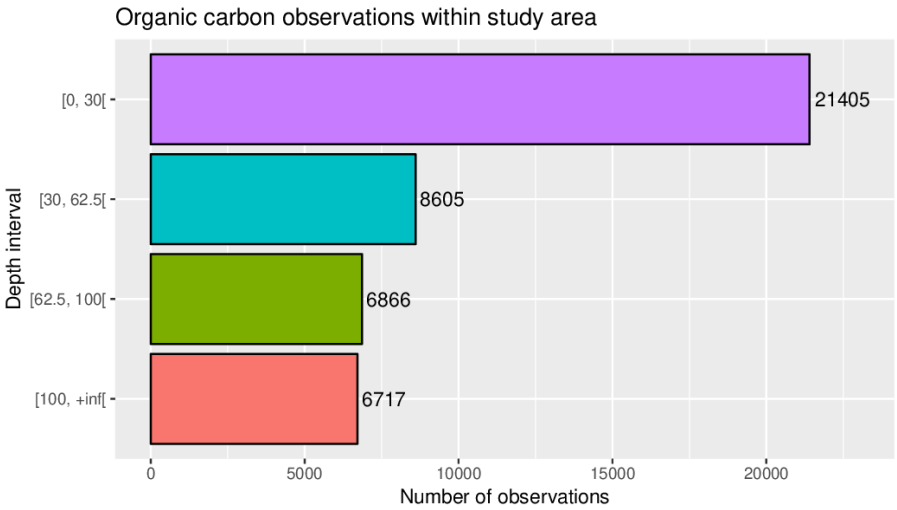

Figure 3.6 *Number of observation by depth interval (defined equidistantly from the four prediction depths).*

Figure 3.6 shows the distribution of observations within four depth intervals defined equidistantly respective to the four prediction depths. About half the number of observations are related to the top 0.3 m layer. However, the number of observations in deeper soil does not decline with depth. For the three soil layers below 0.3 m, the number of observations is rather constant.

3.4 Development of additional covariates

SoilGrids presently uses over 220 different environmental covariates, ranging across various domains: Climate, Ecology, Geology, Land use, terrain Morphology, Vegetation and Water. Considering the current thematic range and the importance of each covariate in previous SOC predictions with SoilGrids, the following additions were appraised:

- Normalised Difference Water Index – NDWI is a vegetation index that provides a proxy for how much water is stored by plants (Gao, 1995). It is an interesting covariate given the importance of the MODIS (Moderate Resolution Imaging Spectroradiometer) near infra-red band (NIR) has had in global SOC predictions performed with SoilGrids;
- JRC products – the Joint Research Centre (JRC) provides various datasets on different topics that are primarily obtained from models applied to earth observation products. Considering the modelled nature of these products, they are unlikely to provide more information to a Machine Learning algorithm than that contained in the original earth observation products;
- Dry matter Productivity – climatic covariates (e.g. temperature, rainfall) reporting to late spring and early autumn have shown high importance in predicting SOC (Hengl et al., 2017). They are possibly proxies for crop growth, which points to the suitability of specialised products like dry matter productivity. A similar alternative, also relevant in this study, is net primary productivity (NPP);
- Soil water – organic matter content is known to correlate positively with soil moisture retention and availability (Fitzpatrick, 1986). A soil water index covariate for the study area was sought for, but no easy-to-use dataset was identified. Instead, datasets reporting surface water extent and change plus the underground water table where used (part of the SoilGrids covariate stack). Some Copernicus products might provide useful additional information but were included in this study, because it would require a substantial effort to prepare these additional covariates. However, for further improvement of the SOM map these products need to be considered.

Based on this analysis the use of NDWI and dry matter productivity covariates was developed further. The NDWI index is computed with the following formula (Gao, 1995):

$$NDWI = (NIR - SWIR) / (NIR + SWIR)$$

Where NIR stands for near infra-red and SWIR for short-wave infra-red. MODIS provides datasets for both wavelengths on a yearly basis. The SoilGrids covariate stack includes these products for the years 2000 and 2014. In fact MODIS provides two different SWIR products: one corresponding to band 7 and another to band 5 (i.e. specific wavelengths) of the Landsat sensor. Therefore, four different covariates were developed for this study:

- **NDWI1L00** – computed with the Landsat band 5 map from 2000;
- **NDWI2L00** – computed with the Landsat band 7 map from 2000;
- **NDWI1L14** – computed with the Landsat band 5 map from 2014;
- **NDWI2L14** – computed with the Landsat band 7 map from 2014.

Since these new covariates were computed from existing covariates, no subsequent spatial processing was necessary.

The MODIS product suite includes a pre-computed edition of NPP maps provided on an yearly basis between 2000 and 2015. The following three products were re-sampled to fit the SoilGrids raster structure and added to the covariate stack:

- **NPP00** – net primary productivity in 2000;
- **NPP15** – net primary productivity in 2015;
- **NPPAVG** – average net primary productivity between 2000 and 2015.

3.5 Prediction of soil organic matter map for the EU

SOC maps were generated with an ensemble model, composed by a weighted average of the outputs from the Random Forests and Gradient Boosting models, which were trained on log-transformed observations of SOC. These are machine-learning methods; more detailed information is given in James et al. (2013). It also allowed the use the 2017 SoilGrids code base in its official form, as available in the public domain. The Random Forests model was trained with 150 decision trees. Beyond that number performance gains are only marginal, at the expense of increased computational time.

Even though the study area is small compared with a global product like SoilGrids, the number of observations and covariates is large enough to demand hours of computation time. Therefore the different computation tasks where parallelised, i.e. coded in order to be executed in multiple CPUs at the same time. Each task is divided in various execution threads, each processed in parallel in a different CPU. With execution parallelised across ten threads, computation times for each task were:

1. Regression matrix creation: 18 minutes;
2. Model fitting (Random Forests and Gradient Boosting): 11 minutes;
3. Prediction (for a single depth): 70 minutes.

Prediction time declines quasi-linearly with an increase of the number of execution threads. Model fitting and the regression matrix creation include steps that cannot be parallelised, therefore computation times do not decline as fast with the increasing number of execution threads.

The resulting predictions are stored as Virtual Raster Tiles. This format is very convenient for parallel processing, with each processing thread taking care of an individual tile. In this study, prediction was parallelised along the tiles described in Section 3.2, with the tiles predicted for a particular depth stored in a specific folder. A VRT meta-data file was created for each depth, referencing the corresponding tiles.

The VRT format is also convenient for coordinate system re-projection. The 2017 SoilGrids source code operates on geographical coordinates, whereas the outputs of this study are delivered in the European system: Lambert's Azimuthal Equal-area projection applied on the ETRF89 ellipsoid positioned on Potsdam. Using the VRT format an actual translation between the two systems is not necessary, a new VRT meta-data file is sufficient. Beyond sparing computation time this scheme also spares disk space.

The final soil organic matter (SOM) maps were obtained by applying a multiplication factor of 1.724 (Fitzpatrick, 1986) to the SOC predictions obtained from the SoilGrids code base. Figure 3.7 presents maps of the SOM predictions for the four depths. SOM declines rapidly with depth in most regions of the EU. Furthermore, the 80 cm and 120 cm predictions are not very different. The largest differences take place between the 15 cm and 45 cm depths, as expected, with much of the SOM present in the top layer of soil.

Uncertainties in the predicted SOM content result in uncertainties in the leaching concentration of pesticides in groundwater. The contribution of uncertainties in this property and substance-specific properties to the uncertainties in the spatial 90^{th} percentile of the leaching concentration in groundwater have been assessed using GeoPEARL by Van den Berg et al. (2012). The parameters that contribute most to the uncertainty in the leaching concentration are the coefficient for sorption on organic matter and the half-life in soil. Soil organic matter was the most important soil property contributing to the uncertainty in the leaching concentration. Therefore, improvement of the reliability and precision of SOM maps would contribute to a better assessment of the risk of leaching to groundwater.

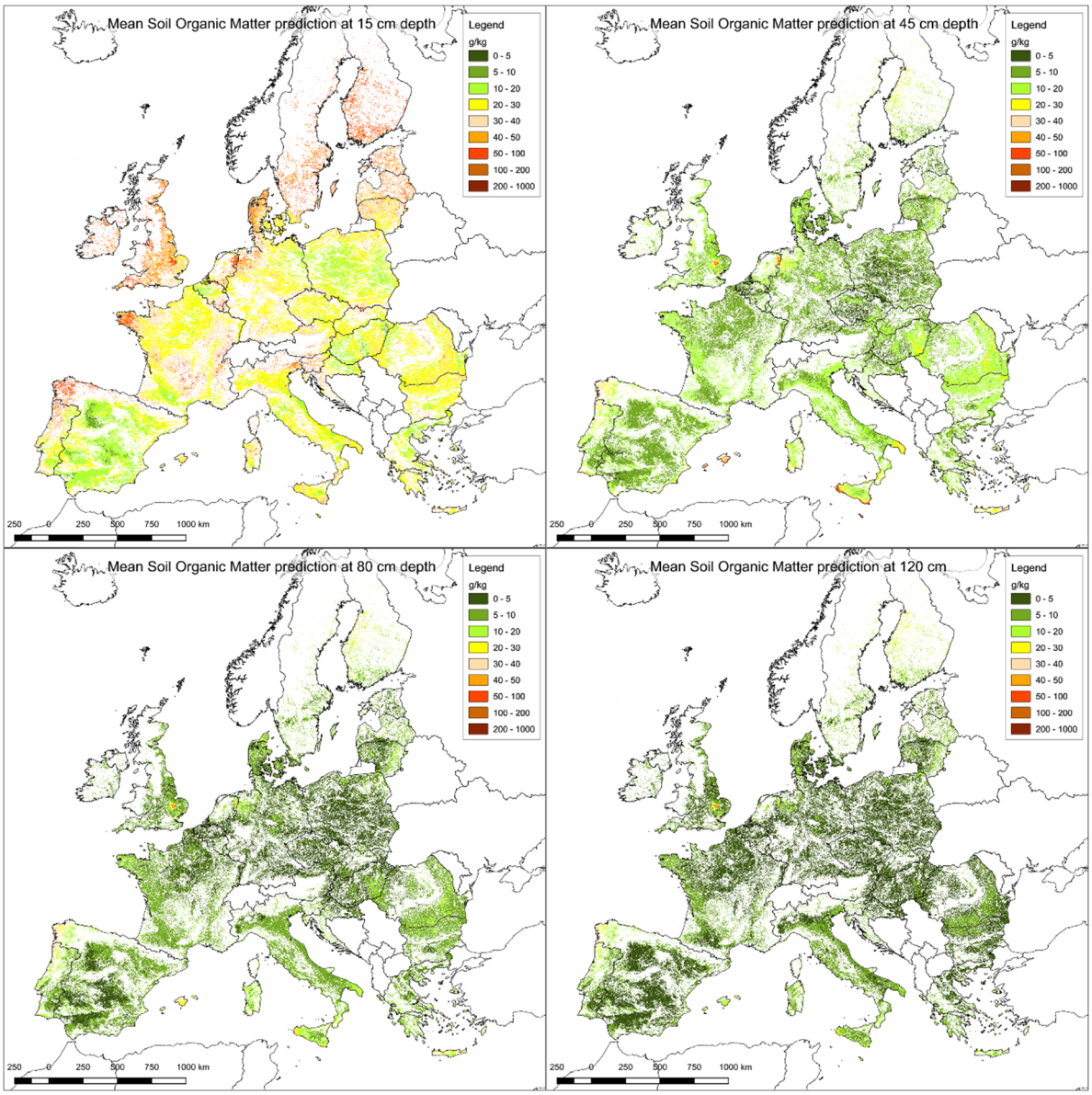

Figure 3.7 *Maps of mean soil organic matter content (g/kg) as predicted by SoilGrids 2018 at four target depths: 0.15 m (top left), 0.45 m (top right), 0.80 m (bottom left) and 1.20 m (bottom right). SoilGrids models trained on log-transformed observations.*

4 Validation procedure for the EU soil organic matter map using SoilGrids

4.1 Cross-validation procedure

A 10-fold cross-validation procedure, adapted from Poggio et al. (2021), was used to assess the performance of the models employed in Chapter 3. This method is usually applied by randomly splitting observations into 10 different sets of similar size. Each set of observations is used once as validation set and is included nine times in the calibration set. In this study applying this 10-fold splitting could lead to spatial bias, since the spatial distribution of soil profiles is rather heterogeneous: dense in some regions and sparse in others (Section 3.3). Therefore, the splitting of observations among validation folds was made by a profile-level spatial stratification, i.e. by dividing the study area into contiguous spatial bins of similar dimension (the strata). The profiles contained within each stratum were then equally distributed among the 10 validation folds, thus guaranteeing that each different region or country was represented in the validation set of each fold.

Considering the size of the study area, a geodetical grid was used to create the spatial strata. A square grid defined on the Cartesian plane would yield large area distortions. The R package *dggrigR* (Barnes, 2016) was used to generate an Icosahedral Snyder Equal-Area Grid (ISEAG) (Sahr et al., 2003), a quasi-regular hexagonal grid covering the entire globe. To each soil profile was assigned a stratum identifier corresponding to the ISEAG cell in which it is located. Table 4.1 synthesises the distribution of profiles among the spatial strata defined by the ISEAG.

Each profile was then assigned to a validation fold with the function *createFolds* from the R package *caret* (Kuhn, 2018). This function implements an algorithm proposed by Hyndman and Athanasopoulos (2013), that aims to balance the relevance of strata with fewer profiles, thus avoiding bias towards dense strata. Figure 4.1 displays the ISEAG cells used as strata and the profiles assigned to two validation folds, demonstrating their similar distribution in geographic space.

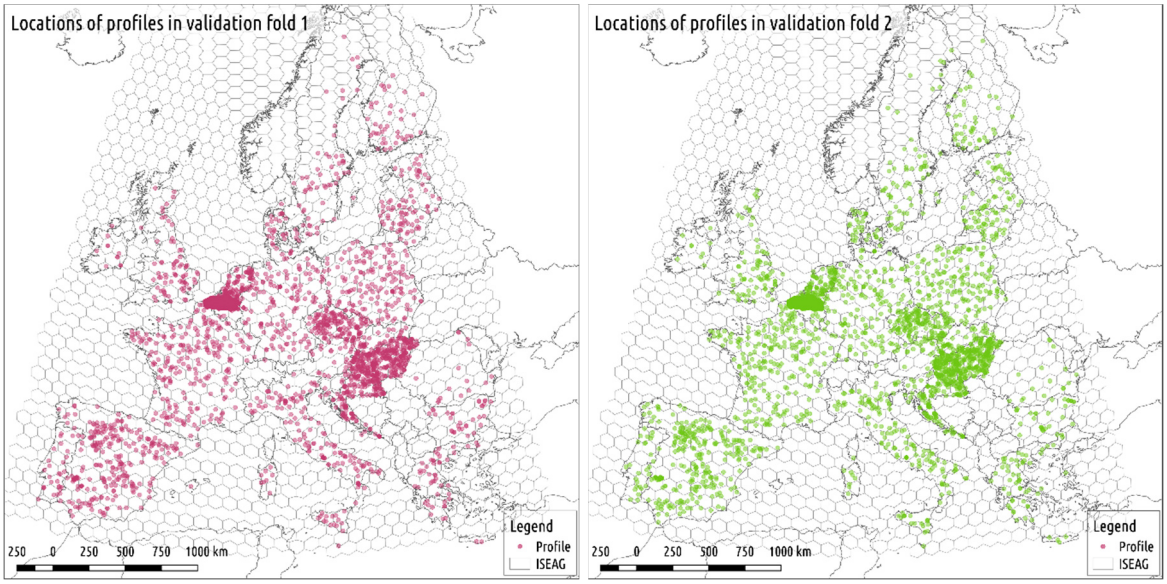

Figure 4.1 *Soil profiles selected for validation folds 1 (left) and 2 (right), overlayed on the ISEAG.*

Table 4.1 *Summary statistics of the distribution of profiles per spatial stratum.*

Min	1st Qu	Median	Mean	3rd Qu.	Max
1.00	5.00	16.00	28.05	32.00	1605.0

4.2 Model performance

Three different models were assessed with the cross-validation procedure described above:

- **Random Forests** (implemented by the *ranger* package);
- **Gradient Boosting** (implemented by the *xgboost* package);
- **Ensemble**: weighted average of gradient boosting and random forests, where the weights are inversely proportional to the MSE of each individual model. This is the same method as described in Hengl et al. (2017, Equation 2).

The metrics applied were Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Error (ME) and Model Efficiency coefficient (MEC). The definition of these metrics has been given in Section 2.1.2.

Table 4.2 gathers performance metrics for the three models assessed. These metrics report to the complete set of observation versus predicted pairs from the 10 validation folds.

Table 4.2 *Performance of model trained on soil organic carbon data with 10-fold cross-validation. Values in table expressed as soil organic matter content.*

Model	RMSE	MAE	ME	MEC
	g/kg	g/kg	g/kg	%
Random Forests (150 trees)	41.9	13.02	-0.47	46
Gradient Boosting	44.8	14.70	-0.30	38
Ensemble [(RF + GB) / 2]	41.7	13.17	-0.39	46

As presented in Table 4.2 the negative values for the mean error (ME) indicate a slight overestimation of the soil organic matter content, but a significant systematic error was not demonstrated. Note also that the mean error is negligible compared to the root mean squared error. The RMSE is much larger than the MAE, so apparently there are some outliers with very large errors. The model efficiency for Random Forests is better than for Gradient Boosting. The result for the Ensemble is very much similar to that for Random Forest alone. The MEC is around 45%, so the models explain about 45% of the variation in SOM.

4.3 Model performance with log-transformed data

4.3.1 Overall performance

In this exercise the original SOC observations were log-transformed before being fed to the models. The expression below was used for transformation:

$$OCL_i = log(OC_i + 1), i = 1, \dots, n$$

in which:
OC_i = organic carbon content (g/kg) observed at site *i*.
OCL_i = log-transformed organic carbon content (g/kg) observed at site *i*.

The models were thus trained on the OCL_i values. To gather model performance metrics the predictions were back-transformed by taking the antilog. Prediction errors (E_i) were computed with the following expression:

$$E_i = OC_i - \left(exp\left(\widehat{OCL_i}\right) - 1\right)$$

in which $\widehat{OCL_i}$ is the log-transformed predicted organic carbon content (g/kg) at observation site *i*.

Table 4.3 *10-fold cross-validation results using a model trained on log-transformed observations of soil organic carbon. Values in table expressed as soil organic matter content.*

Model	RMSE	MAE	ME	MEC
	g/kg	g/kg	g/kg	%
Random Forests (150 trees)	46.8	11.44	5.80	32
Gradient Boosting	50.8	12.44	4.28	20
Ensemble [(RF + GB) / 2]	45.9	11.39	5.49	35

Table 4.3 summarises the 10-fold cross validation results for these models. The performances of the models using log-transformed observations seem to be inferior to those obtained without the log-transformation (see Table 4.2). The model efficiencies decrease from 46% to 35% for the Ensemble model. The mean error values are now positive and differ more from zero than without log-transformation. It should be noted that the over-prediction is caused by the organic matter contents in the higher range, so higher than 10%. There is also a tendency for underestimation (positive ME) by the models using log-transformed observations.

The scatter density plots for the comparison between the observations and the Random Forest, Gradient Boosting and Ensemble model predictions are shown in Figures 4.2 and 4.3. These scatter density plots also show that the scatter is somewhat larger when using the Gradient Boosting model.

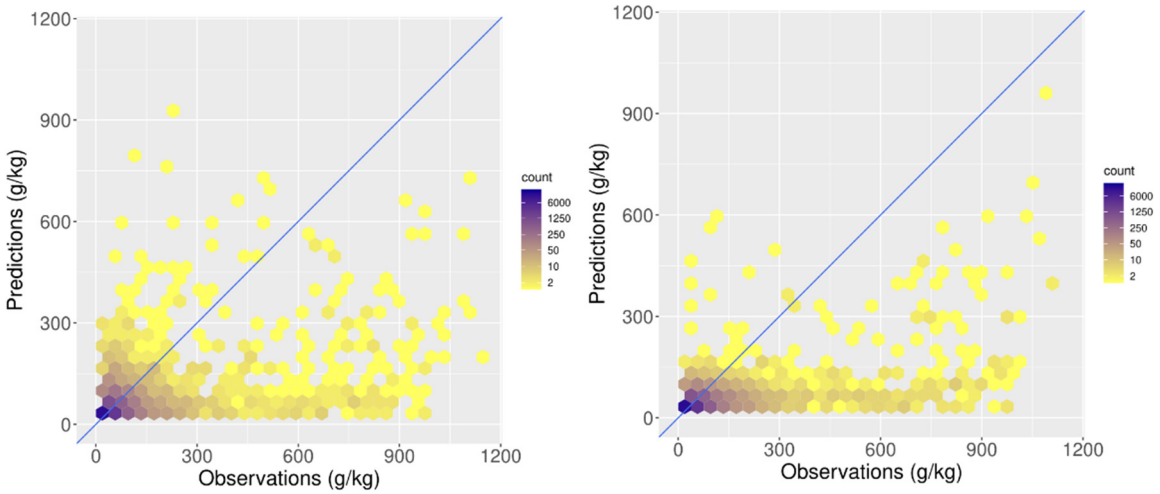

Figure 4.2 *Scatter density plots for comparison of SoilGrids data with the soil organic matter contents computed using the SoilGrids Gradient Boosting model (left) and the SoilGrids Random Forest model (right).*

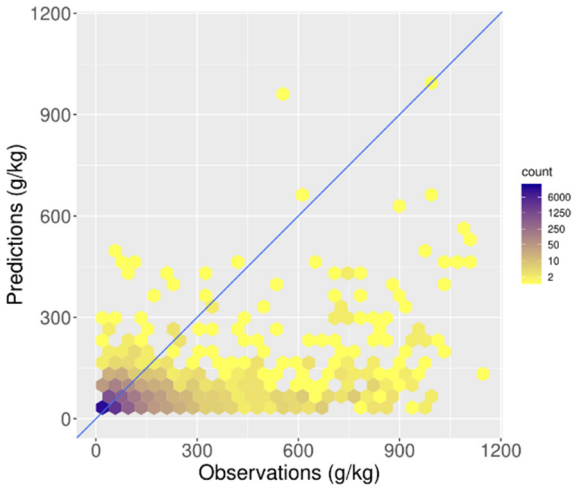

Figure 4.3 *Scatter density plots for comparison of SoilGrids data with the soil organic matter contents computed using the SoilGrids Ensemble model.*

4.3.2 Models performance for SOM below 10%

To evaluate the performance of the models in the range where SOM<10%, the cross-validation approach was applied using all data as described in Section 4.1, but computed the validation metrics on a subset with SOC observations smaller than 5.80% (corresponding with the SOM < 10% case). Table 4.4 reports performance metrics considering only the observations with SOC below 5.8% (58 g/kg, SOM < 100 g/kg). Note that the values presented in this table are SOM contents. For these observations the models trained on the transformed data are superior and also less biased. For these observations the non-transformed models show an over-estimation tendency (negative ME). The negative values for MEC (modelling efficiency coefficient) mean that the prediction is worse than taking simply the average of all observations. Strong negative values, e.g. the value of -108% for the Gradient Boosting model, can be obtained when a model is calibrated on all SOC data and evaluated on a subset. In this case the RMSE is also comparatively large.

Comparing the results of the models trained on the transformed data with those obtained by training the models on all data without log-transform (see Table 4.2), there is a better model efficiency (50 vs. 46) and a lower RMSE value (12.8 vs. 41.7). The mean error for the log-transform on the subset of observations less than 5.8% SOC (less than 10% SOM) is low. Note that these values are obtained after back-transformation, so the mean error of the log-transform values can be expected to be close to zero (Van den Berg et al., 2017).

Table 4.4 *Results of 10-fold cross-validation for observations reporting less than 5.8% soil organic carbon (less than 10% soil organic matter). Values in table expressed as soil organic matter content.*

Log-transform	Model	RMSE	MAE	ME	MEC
		g/kg	g/kg	g/kg	%
No	Random Forests	21.0	9.25	-4.09	-35
No	Gradient Boosting	26.0	10.83	-3.68	-108
No	Ensemble [(RF + GB) / 2]	21.5	9.43	-3.88	-43
Yes	Random Forests	13.2	6.56	0.87	47
Yes	Gradient Boosting	15.5	7.59	0.12	26
Yes	Ensemble [(RF + GB) / 2]	12.8	6.65	0.75	50

The scatter density plots when using the log-transformation for the observations below 5.8% soil organic carbon content (below 10% soil organic matter) are shown in Figure 4.4.

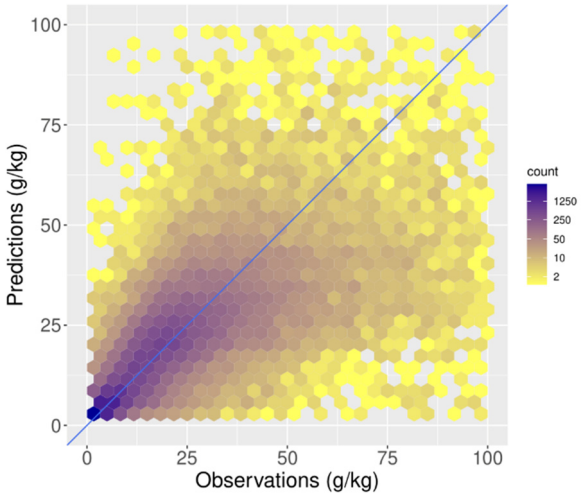

Figure 4.4 *Scatter density plots for comparison of SoilGrids data on soil organic matter contents below 10% (100 g/kg) with those computed using the SoilGrids Ensemble model.*

4.3.3 Models performance for SOM below 5%

To evaluate the performance of the models in the range where SOM<10%, the cross-validation approach was applied using all data as described in Section 4.1, but computed the validation metrics on a subset with SOC observations smaller than 2.90% (corresponding with the SOM < 5% case). Table 4.5 reports performance metrics considering only the observations with SOC below 2.9% (29 g/kg, SOM < 50 g/kg). Note that the values presented in this table are SOM contents. Here the models trained on the non-transformed data present even higher over-estimation tendencies, distancing these further from the models trained on the log-transformed data.

Again, without log-transformation the negative values for MEC mean that the prediction is worse than taking the average of all observations. The negative mean error values mean that the model overestimates the SOC content observed. The best result is obtained for the Ensemble model with log-transformation, with a mean error of soil organic matter of -0.71 g/kg, so 0.07% SOM. The RMSE is also lower for the log-transformed models, and indicate that the predictions of the log-transformed model overall agree fairly well with the SOC observations in the 0-2.9% range (0-5% SOM range).

Table 4.5 *Results on 10-fold cross-validation for observations reporting less than 2.9% soil organic carbon (less than 5% soil organic matter). Values in table expressed as soil organic matter content.*

Log-transform	Model	RMSE	MAE	ME	MEC
		g/kg	g/kg	g/kg	%
No	Random Forests	18.2	7.78	-5.00	-113
No	Gradient Boosting	21.8	9.04	-4.48	-206
No	Ensemble [(RF + GB) / 2]	18.2	7.87	-4.74	-114
Yes	Random Forests	9.90	5.02	-0.73	37
Yes	Gradient Boosting	11.5	5.92	-1.09	15
Yes	Ensemble [(RF + GB) / 2]	9.17	5.11	-0.71	46

The scatter density plots when using the log-transformation for soil organic matter contents below 5% are shown in Figure 4.5.

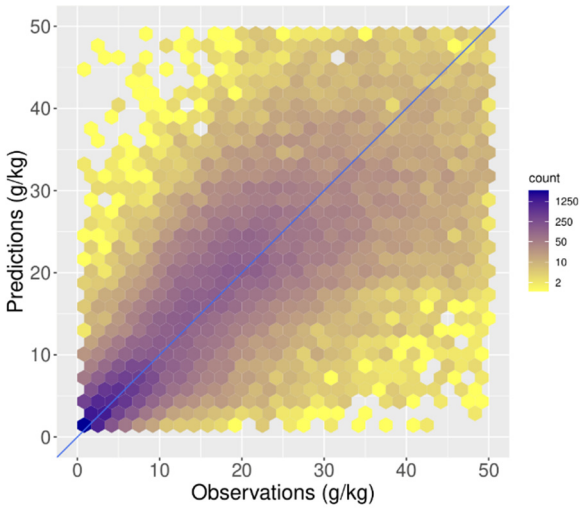

Figure 4.5 *Scatter density plots for comparison of SoilGrids data on soil organic matter contents below 5% with those computed using the SoilGrids Ensemble model.*

4.4 Discussion

The cross-validation results point in the first place to two important conclusions regarding prediction models themselves: (i) Random Forests clearly outperforms Gradient Boosting; and (ii) the Ensemble of models slightly improves over Random Forests. These relations remain true across all datasets addressed, be it with the trimmed sets excluding high SOC observations or with the log-transformed sets. A similar relationship between Gradient Boosting and Random Forests was observed previously with SoilGrids (Hengl et al., 2017). The merit of the ensemble of methods is thus worth pondering upon. In face of computation or time resources constraints, the exclusive employment of Random Forests appears well justified.

The overall performance metrics presented in this study are not as high as those reported for the global SOC model used in 2017 in the SoilGrids project (Hengl et al., 2017, Table 1). However, the cross-validation procedure used in that earlier study is not as sophisticated, since it ignores spatial distribution bias.

The performance of the models used in this study when trained on log-transformed data is considerably poorer than with the original data. Trained on these data, the models show a tendency for under-estimation (positive mean error, see Table 4.3). However, there is a relevant bias caused by high SOC observations. Even though only 2.6% percent of the observations report SOC over 5.8% (SOM > 100 g/kg) and only 8.8% report SOC over 2.9% (SOM > 50 g/kg) these high value observations account for much of the prediction error.

Since this study targeted soils in arable land, where soil organic matter is generally below 5%, it is important that models perform well in this range. The experiments with restricted datasets revealed important weaknesses of these models based on decision trees in face of highly concentrated (and small) observation values. With observations under 5.8% soil organic carbon (under 10% SOM) a heavy tendency for over-estimation is already visible, whereas with observations under 2.9% soil organic carbon (under 5% SOM) model performance can be described as very poor with MEC values < -100 (note that this still equates to more than 90% of the observation set).

It is then for the restricted training sets that the log-transformation shows its benefits. With a log-transformed set of observations under 5.8% soil organic carbon (under 10% SOM) the models yield much better results in this study, particularly Random Forests, with a minimal tendency for under-estimation. With the log-transformed set restricted to under 2.9% soil organic carbon (under 5% SOM), performance of the model was only slightly better than using the 10% SOM threshold. For a detailed discussion on the advantages and disadvantages of modelling in the untransformed and log-transformed space we refer to Lark and Lapworth (2012).

5 Comparison SoilGrids 2018 SOM map with other datasets

5.1 Comparison with EFSA SOM

The final soil organic matter predictions obtained by training the models on all data with log-transform were compared with the soil organic top-soil map published by the European Food Safety Authority (EFSA) (Panagos et al. 2012). This map is based on JRC's topsoil SOC map (JRC 2014). The EFSA map was translated to per mille units (g/kg) and then subtracted from the SoilGrids 2018 SOM map predicted in this study for the 15 cm depth. The result is shown in Figure 5.1.

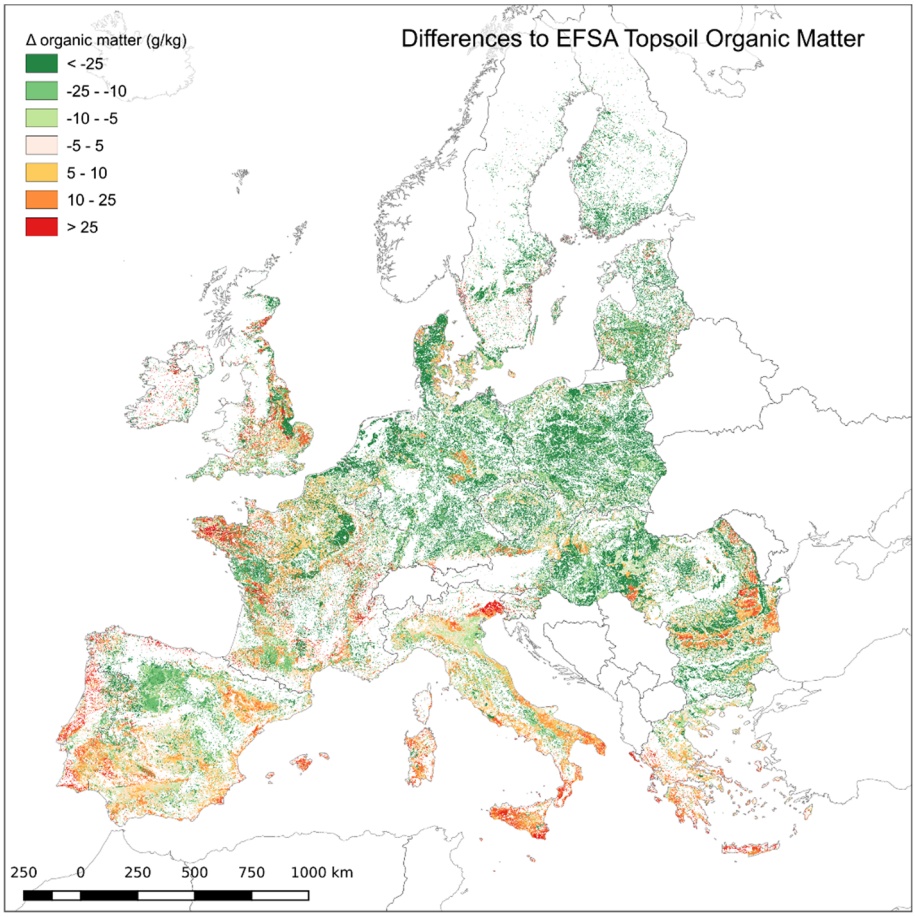

Figure 5.1 *Differences between the SoilGrids 2018 SOM map for arable land in the EU and the EFSA SOM map for the topsoil. Difference negative if SoilGrids SOM 2018 < EFSA SOM.*

In general, the prediction in this study is substantially lower than the EFSA map, in particular for Central Europe. This is an expected result, since the models used in this study where applied only to observations within arable land. SOM tends to be lower in soils subject to this type of land use, therefore, a model trained with observations collected in soils from all land uses may tend to overestimate SOM in arable lands. This seems to be the case with the EFSA top-soil organic matter map, which derives predictions without land use as an explanatory variable for SOM in the EU. Table 5.1 summarises the distribution of differences and Figure 5.2 presents their distribution with a histogram.

Table 5.1 *Distribution of differences between the SoilGrids 2018 SOM and EFSA SOM maps (g/kg).*

Quant. 0.10	Quant. 0.25	Quant. 0.50	Quant. 0.75	Quant. 0.90
-68	-36	-9	4	12

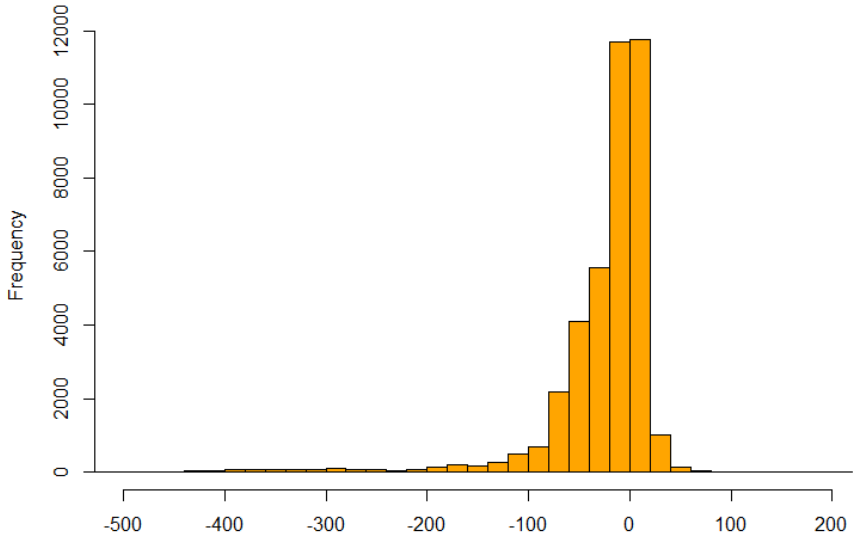

Figure 5.2 *Histogram of differences between the SoilGrids 2018 SOM and EFSA SOM maps (g/kg). Negative values means higher value on the EFSA SOM map. The histogram was calculated on a random sample of 100,000 grid cells from the maps.*

5.2 Comparison with SoilGrids 2017

A similar comparison was conducted between the SoilGrids 2018 SOM map for arable land in the EU and the SoilGrids 2017 SOM map. As before, SOC values were converted to SOM using a conversion factor of 1.724. In this case all four depth intervals could be compared. Since the standard depth levels used in SoilGrids do not match those of this study, the trapezoidal integration rule suggested by Hengl et al. (2017) was applied to derive comparable predictions. This approach is justifiable, since this study focusses on arable land, where soils are predominantly mineral. The SoilGrids 2017 predictions were subtracted from those of this study and the results then projected to the European coordinate system for mapping. Figure 5.3 portraits these comparison results.

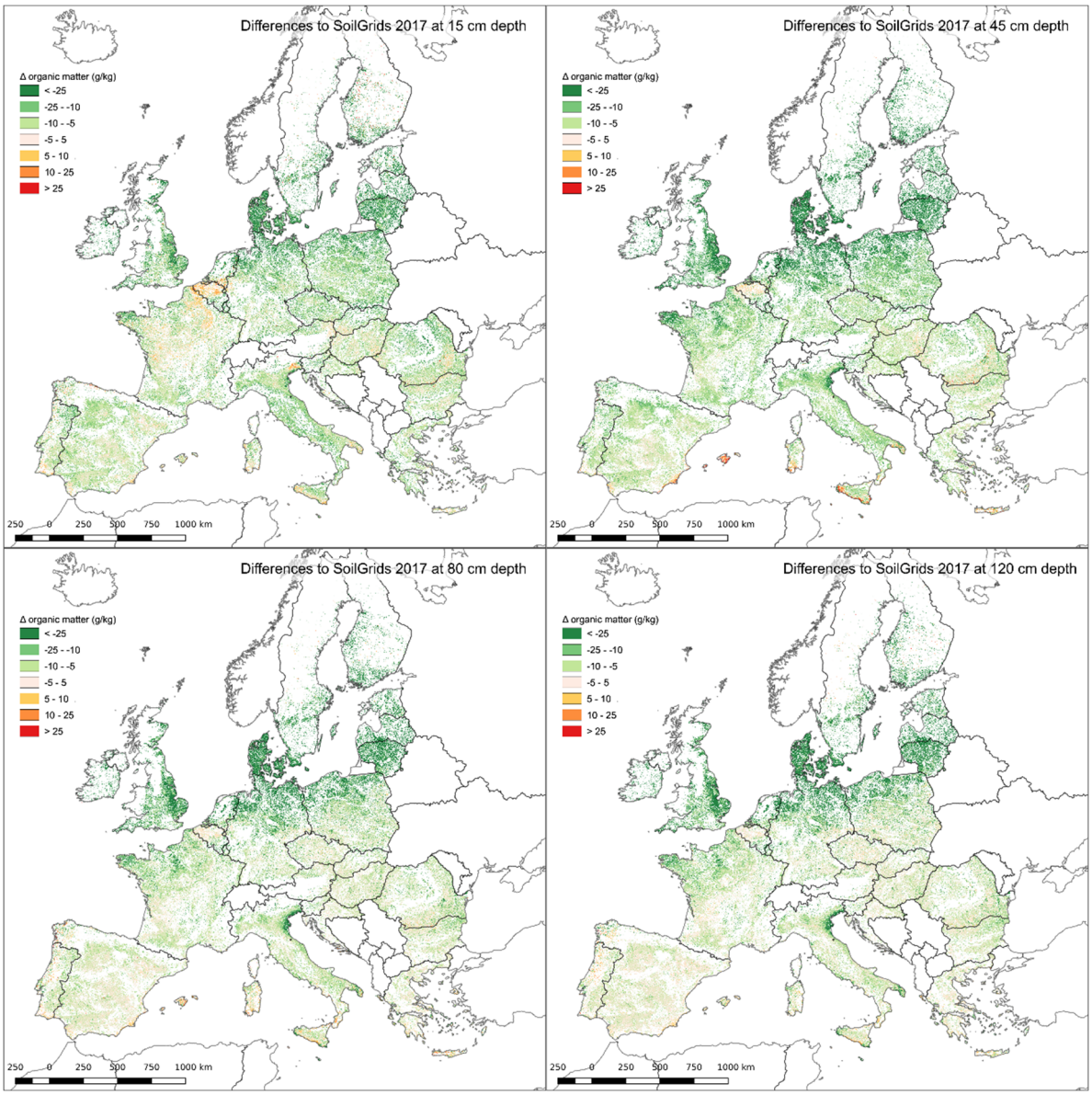

Figure 5.3 *Differences to the SOM predictions for arable land in the EU between SoilGrids 2017 and SoilGrids 2018. Negative values mean overestimation by SoilGrids 2017.*

The predictions in this study are generally lower than those in the SoilGrids 2017 SOM map. The histogram of the differences between the maps based on SoilGrids 2017 and SoilGrids 2018 are shown in Figure 5.4. It is clear that a very skewed distribution is the result of this comparison. The median difference between the two maps for a depth of 0.15 m is about is -4 g/kg, so SoilGrids 2017 has on average larger soil organic matter content predictions than SoilGrids 2018.

The difference between SoilGrids 2017 and SoilGrids 2018 is largely due to the difference in the modelling approaches. The model predictions of SoilGrids 2017 were made based on the observations without the log-transformation to calibrate the model, whereas the SoilGrids 2018 predictions were made after calibration of the log-transformed observations and considering observations on soil organic carbon in arable land only.

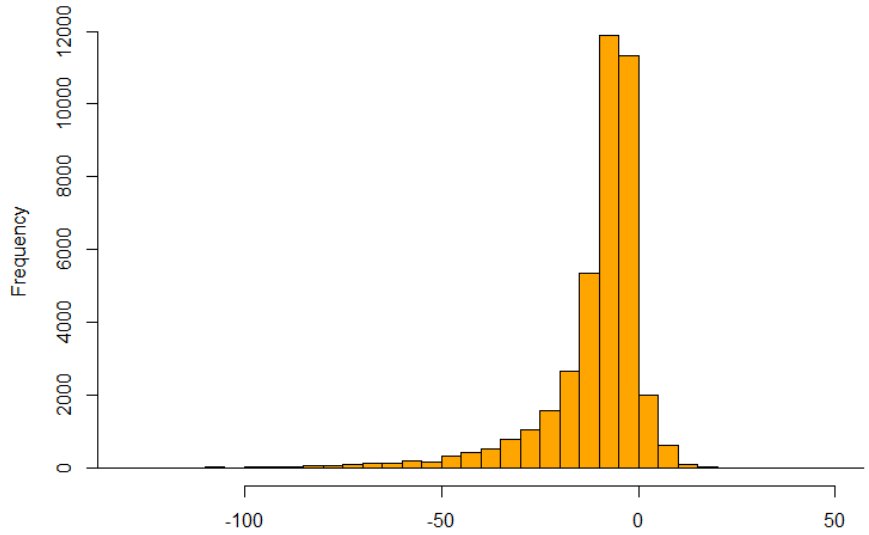

Figure 5.4 *Histograms of SOM (g/kg) differences between SoilGrids 2018 and SoilGrids 2017 at 15 cm depth. Negative values means higher values for SoilGrids 2017. The histogram was calculated on a random sample of 100,000 grid cells from the maps.*

6 Conclusions and Recommendations

6.1 Conclusions

This study started by investigating the performance of SoilGrids 2017 in the Netherlands. SoilGrids 2017 appears on a par with the GeoPEARL SOM map when model predictions are compared with independent observations from the LSK (Landelijke Steekproef Kaarteenheden) data set, albeit with a tendency for overestimation. As it does not log-transform input data, the predictions made by SoilGrids 2017 are naturally biased towards Soil Organic Matter (SOM) rich soils, whereas arable lands have generally low soil organic matter contents in the top soil.

SoilGrids 2017 was further assessed for the whole EU with topsoil SOM observations on arable land collected in the LUCAS project. In comparison with the EFSA 2012 SOM map, SoilGrids 2017 performs considerably better. This may partly be explained by the fact that SoilGrids 2017 used the available LUCAS dataset for calibration of the machine learning model. Here again it should be noted that the SoilGrids 2017 predictions were biased towards SOM-rich soils.

The results of the comparison of SoilGrids 2017 map with the GeoPEARL SOM and EFSA SOM maps showed that it would be worthwhile to adapt the SoilGrids 2017 prediction framework to develop a tailored SOM prediction for the arable lands of the EU.

The tailored SoilGrids 2018 modelling started by assessing the relative performance between three models: Random Forests, Gradient Boosting and the Ensemble. Random Forests proved to be more effective than Gradient Boosting, in some cases with a large margin. The Ensemble of both models yielded an improvement over Random Forests by itself, but the improvement was comparatively small. The relative performance of the models remained the same across the different observation sets tested.

The effect of log-transforming observations on soil organic carbon (SOC) prior to modelling was also investigated in this study. While with original SOC observations models perform better without transformation, when only low SOM contents are considered (< 5% SOM and < 10% SOM) models perform better with log-transformed observations. This is relevant because this study was restricted to arable land, where SOM is mostly below 5%.

Final predictions were therefore carried out with log-transformed observations and the Ensemble model. These predictions showed visible differences to previous SOM products predicted in Europe. Comparison with the EFSA topsoil and the SoilGrids 2017 maps showed that SoilGrids 2018 had systematically lower SOM predictions and resulted in better validation statistics, as demonstrated by negligible bias, a lower root mean square error and a higher model efficiency.

The results of this study show that an improved soil organic matter map for the EU was obtained using SoilGrids 2018. This could be used as input for a soil schematisation for a spatially-distributed model to assess the leaching of pesticides to groundwater at the EU level.

6.2 Recommendations

The results of this study show that SOM predictions within arable land are over-estimated by model approaches that use observations collected in areas with different land use or land cover, such as the SOM map based on the EFSA 2012 dataset. The distribution of SOM observations collected within arable land is substantially lower to that in other areas. This difference is especially visible at the European scale. Therefore, it is recommended to use a tailored approach and appraise only SOM data

from arable land for mapping, if the resulting maps are to be used in applications that operate on arable land only, as is the case for assessing the leaching to groundwater in areas with arable land.

A further important conclusion from this study is the effect of log-transforming observations on model performance. The results obtained here point to the usefulness of log-transforming observations in skewed samples. Models targeting a specific and skewed distribution benefit from this pre-processing step, which is recommended in circumstances like those of this study. However, log-transformation did not improve the prediction of the general SoilGrids model including all available SOM data. For studies in which a wide range of SOM values are of interest it may be more appropriate to abstain from log-transformation prior to modelling.

The SoilGrids SOM map, as obtained in this study, could be used as input for a soil schematisation for a spatially-distributed model to assess the leaching of pesticides to groundwater at the EU level. Such a schematisation is currently being investigated by the SETAC Working Group on Spatially Distributed Leaching Modelling (SETAC-SDLM).

Although SoilGrids 2018 predicts SOM content better than the SOM contents obtained using SoilGrids 2017 or the EFSA 2012 dataset, there is still scope for further improvement. New data on soil organic carbon contents (SOC) in the EU can be added to the SoilGrids training data set. The spatial distribution of observation sites and soils may be taken into account within the framework of the LUCAS project to collect more data on SOC. The improvement may be in particular important in areas with soils for which relatively few data are available. Furthermore, the set of covariates that is used to predict SOC (and consequently SOM) could also be extended and improved, while further advancement of modelling algorithms may also contribute to obtain better predictions.

References

Barnes, R., 2016. dggridR: Discrete Global Grids for R. https://github.com/r-barnes/dggridR

Batjes, N. H., Ribeiro, E., Van Oostrum, A., Leenaars, J., Hengl, T., & de Jesus, J. M., 2017. WoSIS: providing standardised soil profile data for the world. Earth System Science Data, 9(1), 1.

Büttner, G., 2014. CORINE land cover and land cover change products. In Land Use and Land Cover Mapping in Europe (pp. 55-74). Springer, Dordrecht.

Chen, J., Chen, J., Liao, A., Cao X., Chen L., Chen X., He C., Han G., Peng S., Lu M., Zhang W., Tong X, & Mills, J. 2015, Global land cover mapping at 30 m resolution: A POK-based operational approach. ISPRS Journal of Photogrammetry and Remote Sensing 103, 7-27. doi: 10.1016/j.isprsjprs.2014.09.002.

European Commission, 2014. Assessing Potential for Movement of Active Substances and their Metabolites to Ground Water in the EU. Report of the FOCUS Ground Water Work Group, EC Document Reference Sanco/13144/2010 version 3, 613 pp.

EC, 2012. EFSA Spatial Data Version 1.1 Data Properties and Processing, JRC Technical report EUR 25546 EN.

EFSA, 2015. European Food Safety Authority (EFSA) Data & PERSAM software tool https://esdac.jrc.ec.europa.eu/content/european-food-safety-authority-efsa-data-persam-software-tool

EFSA, 2017. EFSA Guidance Document for predicting environmental concentrations of active substances of plant protection products and transformation products of these active substances in soil. EFSA Journal 2017, 15(10), 4982.[115 pp.] doi: https://doi.org/10.2903/j.efsa.2017.4982

Finke, P.A., De Gruijter, J.J. and Visschers, R., 2001. Status 2001 Landelijke Steekproef Kaarteenheden en toepassingen. Rapport 389, Alterra Wageningen UR.

Fitzpatrick, E. A., 1986. An Introduction to Soil Science (Second Edition). Longman Scientific & Technical.

Gao, B. C., 1995. Normalized difference water index for remote sensing of vegetation liquid water from space. In Imaging Spectrometry (Vol. 2480, pp. 225-237). International Society for Optics and Photonics.

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Oxford University Press, New York.

Hengl, T., Mendes de Jesus, J., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A. Kempen, B., Leenaars, J.G.B., Walsh, M.G. & Ruiperez Gonzalez, M., 2014. SoilGrids1km - Global Soil Information Based on Automated Mapping, PLoS one, 9(8): e105992. doi:10.1371/journal.pone.0105992

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X, Bauer-Marschallinger, B, Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel. S., & Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS one, 12(2), e0169748.

Hyndman, R.J. and Athanasopoulos, G., 2013. Forecasting: principles and practice. https://www.otexts.org/fpp

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An Introduction to Statistical Learning. With Applications in R., Springer.

JRC, 2014. New European map of topsoil organic carbon. https://ec.europa.eu/jrc/en/science-update/european-map-topsoil-organic-carbon

Kuhn M. Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M. Lescarbeau, R., Ziem, A., Scrucca, L., Tang Y. Candan, C., & Hunt, T., 2018. Package 'caret'. https://cran.r-project.org/web/packages/caret/caret.pdf

Lark, R.M., & Lapworth, D.J., 2012. Quality measures for soil surveys by lognormal kriging. Geoderma 173-174, 231-240.

Nash, J.E., & Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part I – A discussion of principles. Journal of Hydrology 10, 282-290.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., & Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. European Journal of Soil Science, 69(1), 140-153.

Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L., 2012.European Soil Data Centre: Response to European policy support and public data requirements. Land Use Policy 29, 329-338.

Poggio, L., de Sousa, L.M., Batjes N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E. and Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL, 7, 217-240. https://doi.org/10.5194/soil-7-217-2021

Reijneveld, A., Van Wensem, J., and Oenema, O., 2009. Soil organic carbon contents of agricultural land in the Netherlands between 1984 and 2004, Geoderma 152, 231-238.

Sahr, K., White, D., & Kimerling, A. J., 2003. Geodesic discrete global grid systems. Cartography and Geographic Information Science, 30(2), 121-134.

Sleutel, S., De Neve, S., Singier, B. and G. Hofman, 2007. Quantification of Organic Carbon in Soils: A Comparison of Methodologies and Assessment of the Carbon Content of Organic Matter. Communications in Soil Science and Plant Analysis, 38, 2647 -2657.

Tiktak, A., Van der Linden, A.M.A. & Boesten, J.J.T.I., 2003. The GeoPEARL model. Model description, applications and manual. RIVM Report 716601007.

Berg, F. van den, Tiktak, A., Heuvelink, G.B.M, Burgers, S.L.G.E., Brus, D., De Vries, F., Stolte, J. and J.G. Kroes, 2012. Propagation of uncertainties in soil and pesticide properties to pesticide leaching, Journal of Environmental Quality, 41, 253-261.

Van den Berg, F., Tiktak, A., Hoogland, T., Poot, A., Boesten, J.J.T.I., Van der Linden, A.M.A., Pol, J.W. 2017. An improved soil organic matter map for GeoPEARL_NL. Wageningen University and Research. Report 2816, Wageningen, the Netherlands.

Van der Linden, A.M.A., Boesten, J.J.T.I., Cornelese, A.A., Kruijne, R., Leistra, M., Linders, J.B.H.J, Pol, J.W., Tiktak, A. and A.J. Verschoor, The new decision tree for the evaluation of pesticide leaching from soils., RIVM report 601450019/2004.

Visschers, R., Finke, P.A. and De Gruijter, J.J. 2007. A soil sampling program for the Netherlands. Geoderma, 139, 60-72.

Wageningen Environmental Research
P.O. Box 47
6700 AA Wageningen
The Netherlands
T +31 (0)317 48 07 00
www.wur.nl/environmental-research

Wageningen Environmental Research
Report 3126
ISSN 1566-7197

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 6,800 employees (6,000 fte) and 12,900 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.

To explore the potential of nature to improve the quality of life

Wageningen Environmental Research
P.O. Box 47
6700 AB Wageningen
The Netherlands
T +31 (0) 317 48 07 00
www.wur.eu/environmental-research

Report 3126
ISSN 1566-7197

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 6,800 employees (6,000 fte) and 12,900 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.