

Enhancing European soil monitoring with citizen science data

Céline Laurent

`celine.laurent@unibz.it`

Competence Centre for Plant Health, Free University of Bolzano

Claudia Cappello

Free University of Bolzano

Timo Breure

Joint Research Centre

Oleg Osychenko

Quanta Systems S.L.

Arwyn Jones

European Commission, Joint Research Centre

Cristiano Ballabio

Joint Research Centre - European Commission <https://orcid.org/0000-0001-7452-9271>

Panos Panagos

European Commission, Joint Research Centre <https://orcid.org/0000-0003-1484-2738>

Tanja Mimmo

Free University of Bozen-Bolzano

Article

Keywords: environmental monitoring, EU Mission Soil, participatory approach, open data, soil health, soil mapping

Posted Date: August 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10276372/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Title: Enhancing European soil monitoring with citizen science data

Author information: Céline Laurent^{1*}, Claudia Cappello¹, Timo Breure², Oleg Osychenko³, Arwyn Jones², Cristiano Ballabio², Panos Panagos², Tanja Mimmo^{1,4}

¹Competence Centre for Plant Health, Free University of Bolzano, Bolzano 39100, Italy

²European Commission, Joint Research Centre (JRC), Ispra, Italy

³Quanta Systems S.L., Badalona 08912, Spain

⁴Faculty of Agricultural, Environmental and Food Sciences, Free University of Bolzano, Bolzano 39100, Italy

*Corresponding author

Abstract

Soil degradation affects 60–70% of soils in Europe, highlighting the need for large-scale monitoring. Existing programmes, such as the LUCAS Soil Module, provide harmonised datasets but have limited land-use coverage, while integrating national monitoring is constrained by differences in indicators and methods. Citizen science (CS) could extend monitoring coverage, but its data quality and comparability require evaluation. Here, we assessed CS measurements by comparing them with laboratory measurements and LUCAS Soil 2018 data for soil pH, soil organic carbon (SOC) and texture. CS pH measurements captured broad spatial patterns, whereas SOC and texture were more suitable as qualitative indicators. When integrated with LUCAS Soil, CS pH data improved digital soil mapping by reducing prediction error and improving prediction uncertainty. CS data also increased coverage of underrepresented environments, particularly urban soils. Together, our results show that citizen science can complement European soil monitoring when integrated appropriately.

Keywords: environmental monitoring, EU Mission Soil, participatory approach, open data, soil health, soil mapping

Introduction

The degradation of soil has emerged as a major environmental challenge in Europe. According to the EU Soil Observatory Dashboard, 60-70% of soils in the European Union are considered as degraded highlighting the urgent need for consistent monitoring to prevent soil from further degradation¹. The Soil Monitoring and Resilience Directive² establishes a framework at the EU scale to assess, monitor, and improve soil health. Soil

monitoring is essential for assessing soil health as it enables the tracking of various soil physical, chemical and biological properties over time revealing changes in soil properties and functions. It also provides valuable data to guide policy development, inform decision-making for soil health preservation, and to support soil conservation efforts from local to global scale³. The European Commission started the Land Use/Land Cover Area Frame survey (LUCAS) Soil in 2009, a soil monitoring programme dedicated to the collection and analyses of topsoil samples across European Union member states, resulting in a harmonised dataset at European scale, with agricultural soils mainly represented. Many EU member states operate national soil monitoring that could, in principle, complement LUCAS Soil and increase monitoring density. However, several studies have underlined the challenges of comparing soil data across countries^{4–6} highlighting the need for better harmonisation (i.e. unified use of indicators) and standardisation of sampling and analysis between countries. The comparison between soil national monitoring data from European countries and LUCAS Soil remains also difficult with substantial differences in measuring key soil properties (e.g. pH, soil organic carbon - SOC) when analysed at national level due to different sampling strategies (e.g. land use/cover distribution) and analytical methods⁷. To achieve harmonisation of data and enable direct comparison, the LUCAS 2022 survey implemented an intercomparison between LUCAS approach and 13 member states laboratories procedures with the aim to develop pedotransfer functions¹. To support the standardization of methods in soil health assessment over time at European scale, laboratory methods remain the reference standard as they are accurate and reproducible, but they are also resource-intensive, require important laboratory infrastructure and trained professionals to carry such analysis at this broad scale. Beyond harmonisation, LUCAS Soil was principally designed to deliver representative areal statistics at EU and regional scale⁸, rather than fine-resolution spatial information on soil properties. This constrains its direct use for applications such as digital soil mapping, which typically benefits from integrating complementary data sources such as Earth Observation data alongside legacy and survey data to improve spatial and temporal coverage (refs).

Citizen science (CS) has emerged as one of such complementary approach to address these limitations and to broaden monitoring coverage. Soil health knowledge derives mainly from agricultural context and urban areas area often underrepresented⁹. Thus, engaging citizens in standardised field sampling could reach private gardens, urban green areas and other sites that national monitoring programmes often miss. Mason et al. (2024)¹⁰ reviewed 24 European CS projects on agricultural soils and found that over 66% provided soil biodiversity data, while 54% and 42% provided data on vegetation cover and SOC, respectively, demonstrating that soil CS projects represent a potential resource for European soil health assessment and monitoring. In addition, Charvat et al. (2025)¹¹ demonstrated that CS can serve as a valuable approach to enhance spatial and temporal

scale of soil monitoring. Nevertheless, they highlighted that challenges remain in CS projects with variation in data validation rigour and stakeholder engagement that led to inconsistent datasets or difficulties in comparing CS-generated datasets. Overall, the integration of CS-generated soil data to scientific soil monitoring programmes raises important questions about the data quality. The comparison of CS measurements against reference datasets is therefore a crucial step before such data can be implemented into soil monitoring programmes or soil mapping.

Using CS data collected through the project ECHO¹², we assessed the potential of CS measurements to complement the existing LUCAS survey for soil monitoring at the European scale focusing on pH, SOC and texture data. Specifically, we: (i) evaluated the quality of CS data collected within the framework of ECHO against reference data measured by standard protocols; (ii) quantified the agreement, deviation and bias between CS measurements and LUCAS 2018 dataset; and (iii) tested whether integrating CS measurements with the LUCAS 2018 improves digital soil mapping.

Results

Citizen science methods versus laboratory-based methods

Soil pH

The soil pH assessed by citizen scientists in the field ranged from 4 to 9 while the soil pH measured by laboratory method ranged from 4.7 to 8.4, with median values of 7 and 7.7, respectively (Fig. 1A). The scatter plot showed that citizen scientists pH measurements, clustered along distinct horizontal bands, due to the 0.5 pH unit resolution of the citizen scientists' method. The comparison between CS and laboratory-based pH measurements showed an overall moderate agreement ($\text{adj}R^2 = 0.46$), explaining 46% of the variance in the pH laboratory values. The RMSE of 1 pH unit ($\text{rRMSE} = 13.6\%$) shows a moderate deviation from CS measurements compared to laboratory-based method, supporting the overall suitability of the method. However, an underestimation of pH values by citizen scientists was observed, with pH values on average 0.6 lower ($\text{rMBE} = -7.8\%$) compared to the laboratory-based pH measurements.

Soil organic carbon content

The soil organic carbon (SOC) content assessed by citizen scientists in the field ranged from 1.5 to 5 % while the SOC content measured using laboratory-based methods ranged from 0.3 to 18.8 %, with median values of 2 and 2.2 %, respectively (Fig. 1B). The comparison between CS and laboratory-based measurement showed poor agreement

(adjR² = 0.13), with an RMSE of 3.92% SOC (rRMSE = 106%), indicating substantial deviation of CS assessment compared to the laboratory-based measurement. At SOC content above 5%, a systematic underestimation was observed, as showed by the overall MBE of -1.44% SOC (rMBE = -39%), with a plateau between 2-3.5% SOC estimated by citizen scientists, regardless of the laboratory-based SOC content. When restricting the comparison to SOC content below 5% (Fig. S1), the reliability remained poor (adjR² < 0,1), although the magnitude of the deviation was reduced (RMSE = 0.99% and rRMSE = 47.4%) and no systematic underestimation was detected (MBE = 0.03% SOC and rMBE = 1.34%). Accordingly, these results indicated that the colour-chart-based field method used to estimate SOC content was not a reliable approach, particularly for SOC content above 5%.

Soil texture

The confusion matrix showed a poor overall accuracy of 22% and near-zero weighted Kappa coefficient (0.071) indicating poor agreement between CS and laboratory-based soil texture measurement. Only 22% of the samples were assess correctly (one clay loam, two sand and nineteen sandy loam soils) (Table 1). Based on laboratory-based soil texture measurements, most of the soils fell in the sandy loam texture (80%), whereas citizen scientists misclassified these soils across the other texture classes. Despite the poor overall accuracy, the confusion matrix revealed some sensitivity to soil texture. On average, in 25% of the cases, clay loam and clay classes were laboratory-confirmed as clay-related textures. In contrast, soil citizen scientists classified as sand, sandy-clay or sandy showed stronger agreement with sandy-related textures (90%). Notably, silt-related classes showed no correct classification: soil citizen scientists classified as silty-related texture were laboratory-confirmed as sandy loam in on average 78% of cases. Overall, these results suggest that citizens showed poor ability to detect texture classes except for detecting coarse fractions. The absence of laboratory-confirmed silty-related texture and clay texture also limit our ability to conclude on sensitivity to specific texture classes.

Comparison of ECHOREPO citizen scientists' data with LUCAS

Land cover type differences between ECHO samples and LUCAS 2018

The distance between the nearest LUCAS points and ECHOREOPO sampling sites ranged from 87.5 m to 220 km, with a median value of 8 km. The distribution of land cover types at ECHOREPO sampling sites differed from that of the nearest LUCAS points. Urban areas represented the predominant type of land cover type at ECHOREPO sampling sites (31.1%) (Fig. 2A), while arable land was the predominant type of the nearest LUCAS points (37.7%) (Fig. 2B). Forest and heterogeneous agricultural areas were present in both

datasets, with some slight differences in the proportions. Urban areas were mainly underrepresented in the nearest LUCAS points (4.9%) and land cover types such as pastures and permanent crops well represented in the nearest LUCAS points were underrepresented at ECHOREPO sampling sites, with 4.9% and 5.5%, respectively. The correspondence between land cover types at ECHOREPO sampling sites and those at the nearest LUCAS point varied across land cover types (Fig. S2). Arable land showed the highest level of agreement with 62.5% of ECHOREPO arable land associated with arable land at the nearest LUCAS points, followed by permanent crops (43.3%), pastures (36%), forests (30%) and shrubs and herbaceous vegetation (23.4%). Urban areas showed the weakest agreement (9.6%) and no correspondence was found between ECHOREPO and LUCAS sites for inland wetlands, inland waters, and marine waters, while coastal wetland were exclusively present at ECHOREPO sampling sites.

Comparison of citizen scientists data and LUCAS-derived soil pH raster

Figure 3 shows the pH values from ECHOREPO as a function of pH derived from LUCAS raster for across 22 European countries ($n = 6602$). The pH values measured by citizen scientists ranged from 3 to 9, while the soil pH values derived from LUCAS raster data ranged from 4.2 to 8.3, with medians of 6 and 6.4 pH unit, respectively (Fig. 3). CS pH measurements showed a moderate deviation from LUCAS-derived raster pH values ($\text{adj}R^2 = 0.28$, $\text{RMSE} = 1.28$, $\text{rRMSE} = 19.86\%$). CS pH measurements showed a systematic underestimation of pH values by 0.39 pH units on average ($\text{rMBE} = -6.13\%$), compared to LUCAS-derived raster pH values. The RMSE for pH at NUTS2 level showed regional variability, where low RMSE values were mainly found in parts of central and eastern Europe, including some parts of Germany and Poland, and isolated spots in Spain, while higher RMSE values were found in some parts of Germany, Italy and Spain (Fig. 4A). At the NUTS2 level, the median pH measured by CS showed a lower deviation from the LUCAS-derived raster median pH ($\text{adj}R^2 = 0.49$, $\text{RMSE} = 0.85$ pH unit, $\text{rRMSE} = 13.3\%$) compared to the results obtained at the individual measurement level (Fig. 3). A systematic underestimation of citizen scientists' median pH was also observed at NUTS2 level ($\text{MBE} = -0.39$ pH units, $\text{rMBE} = -6.14\%$) (Fig. 4B). Additionally, at the NUTS2 level, RMSE on pH was not significantly related to sampling density (Spearman's $\rho = 0.09$, $p\text{-value} > 0.05$; Fig S3), suggesting that denser sampling did not reliably improve agreement at this broader-scale.

Comparison of citizen scientists and LUCAS-derived raster soil texture across Europe

The confusion matrix showed a low overall accuracy of 24.6% and a near-zero weighted Kappa coefficient (< 0.1) (Table 2), indicating poor agreement between soil texture assessments by the citizen scientists and the LUCAS-derived raster soil texture classes. Sandy loam was the most represented texture class and was showing the highest percentage of correct matches (82%). By contrast, silty clay and clayey soil texture classes showed no correct classification, with both 0% of correct matches. Sandy-related texture classes were correctly assigned by citizen scientists in an average of 82% of cases to sandy-related texture classes, while silty- and clayey-related texture classes were correctly assigned to their fraction-related texture classes in only 9% and 15% of cases, respectively. These results suggest that at a broader spatial scale, citizen scientists reliably detected only the coarser soil fraction. The Kappa coefficient was computed at the NUTS2 level, with predominantly near-zero values, highlighting the poor agreement at this spatial scale (Fig. 5A). Three regions of Germany and one of Italy showed Kappa coefficient between 0.2 and 0.4 (Fig. 5A), representing fair agreement. Only one region of Germany and one region of Belgium showed Kappa coefficient above 0.4, which represent moderate agreement. The relationship between Kappa coefficient at NUTS2 level and the sampling density didn't show any specific trend, the regression line with $\text{adj}R^2 < 0.1$ indicating that the poor agreement is not imputable to the sampling effort (Fig. 5B). The extreme Kappa coefficient was systematically associated with low sampling size, as indicated by low point size. Accordingly, these results showed a poor agreement between the two datasets and an increasing disagreement in texture classes from coarse to fine soil fractions.

Integrating CS measurements with LUCAS through digital soil mapping

The covariates with the highest median importance across the 100 iterations were the aridity index, `mir_PCAb1`, the geographic coordinates and `red_PCAb2`, which were also selected in each iteration (Fig. 6A). The feature selection procedure seemed relatively stable, where 17 out of 24 selected covariates were selected in more than half of the iterations (Fig. 6A).

For the prediction metrics on the hold-out test set, there was a clear distinction between when the CS data was considered to have equal information content (Eq.), and when weighting was applied (Fig. 6B). That is, the error, R^2 and IS metrics degraded with higher sample sizes for Eq. ($N > 1k$), whereas it remained stable up till 3-4k for when weighting was applied. An exception was the PI-Q50 metric, that showed the same trend as Eq., and to a lesser degree when PI-Q50 was combined with the IS weighting (Fig. 6B).

Weighing the CS data based on prediction interval metrics led to better overall performance compared to Equal. However, in general after when $N > 2k$, notably the R^2

metric started to decline (Fig. 6B). Between the different weighting schemes, were relatively small differences occurred (Fig. 6B). These differences consisted mainly of: i.) more informative prediction intervals (low IS and median PICP of 0.90) across sample sizes, ii.) the degree to which a lower error rate compared to LUCAS was maintained without a decreasing R^2 , i.e. maintaining the bias-variance trade-off across larger sample sizes. Overall, the IS-PIQ50-PIC metric performed the best of all weighting metrics.

Whereas the addition of CS measurements had a modest effect compared to LUCAS on the performance for the test set, larger differences occurred when prediction metrics were calculated for each NUTS2 region (Fig. 7A). We calculated the metrics for each NUTS2 region for the IS-PIQ50-PIC implementation, for a sample size ($N=2k$). Comparing these to the LUCAS implementation, showed that 73 out of 138 NUTS2 regions were significantly different (WRST statistic, $p < 0.05$). The majority of these showed a positive difference (40 out of 73), meaning that the CS implementation performed worse compared to LUCAS (Fig. 7A). No specific latitudinal- or longitudinal trends were found for the mean direction of the differences, or the WSRT significance (Fig. 7B). However, the frequency in which a CS measurement was selected across the 100 iterations showed some degree of spatial clustering, particularly the samples that were selected in each iteration (pink circles, Fig. 7C).

Discussion

The assessment of data quality in CS remains challenging, as CS-generated data have been shown to vary depending on citizen scientists experience, task complexity and the use of standardised protocols (Kosmala et al 2016). Although pH, SOC content are among the most frequently monitored indicators in CS projects on soil¹¹, the reliability of the results is context dependent and cannot be assumed to meet the laboratory-based measurements. In our study, the overall quality and agreement of CS soil assessment was strongly dependent on the soil property considered. The limitations reflect inherent constraints of the methods but suggest also that citizen science-generated data require carefully evaluation and when appropriate, bias correction procedure or spatial aggregation of data could be applied for broad-scale soil monitoring.

The poor agreement between CS and laboratory-based SOC measurements (Fig. 1B) indicates limitations of the soil colorimetric assessment. The CS SOC estimated values were restricted in a narrow range between 1.5 to 5% generating a plateau between 2 and 3.5%, while laboratory-based method measurements ranged up to 18.8% (Fig 1B). This limited range resulted in a systematic underestimation of SOC content above 5% and was unable to reflect the variability of SOC content across ECHOREPO samples. This constraint appears to arise from the restricted SOC range in the colour chart rather than

from an inability of citizen scientists to discriminate soil colour. Ferrando Jorge et al. (2021)¹³, applying a Munsell-based colour approach with citizen scientists, used spectrally calibrated soil-colour observations and showed that calibration improved SOC prediction. However, our results indicate that this SOC value range is insufficient to represent the full variability of SOC content. The colour chart does not encompass the diversity of SOC contents observed across Europe, where topsoil organic carbon maps reveal considerable spatial variability, including SOC contents exceeding 5% in several Northern European countries¹⁴.

When the comparison was restricted to the range of SOC content from 1.5 to 5% (Fig. S1), the magnitude of the deviation decreased but the agreement remained poor between citizen SOC assessments and laboratory-based SOC measurements. This suggests that our colour chart method lacks precision and the method is not accurate for quantitative assessment of SOC. Furthermore, soil colour can be influenced by confounding factors such as soil moisture, particle size, mineral composition or the presence of iron oxides¹⁵⁻¹⁷ and lighting effect itself.

Overall, these results suggest that the current colour-chart-based approach is not suitable for quantitative assessment of SOC content. The limitations observed for the colour-chart-based SOC assessment also point to a possible direction for methodological improvement. Within ECHO, we have developed a prototype algorithm for the automatic estimation of soil organic matter based on computer vision analysis of soil sample images¹⁸. The approach uses photographs of the collected soil sample, colour adjustment procedures to reduce the influence of variable lighting conditions and device-specific colour rendering, and the Munsell Soil Colour Charts. The aim is not simply to reproduce the current visual colour-chart method digitally, but to convert image-derived colour information into a more standardised and reproducible representation that can be calibrated against laboratory-based measurements. The prototype is currently in the testing phase, and further validation against laboratory-based organic carbon measurements will be required before it can be used as a quantitative tool. Nevertheless, this development illustrates how CS protocols may be improved by combining simple field observations with colour-standardised image analysis, automated correction procedures and calibration against reference datasets.

Concerning texture, the poor overall accuracy and the near-zero Kappa coefficient indicated that the texture-by-feel method doesn't provide reliable texture classification, across the simplified range adapted from USDA texture classes. Nevertheless, the confusion matrix showed not completely random misclassification by citizen scientists for sandy-related texture classes (Table 1, Table 2). The accuracy of the texture-by-feel method is dependent on local knowledge of the pedological and mineralogical context and is influenced by soil properties such as the soil organic matter, which affects the

tactile characteristics of the soil¹⁹. Salley et al. (2018)²⁰ found that observer experience mattered: professional soil scientists (academically-trained) correctly classified 66% of samples, compared with 27-41% for seasonal field technicians (moderately-trained). The accuracy was increased to 91% and to 71-78%, respectively, when adjacent texture classes were accepted as correct. The authors therefore argued that improving soil texture estimation requires training, calibration and decision-support tools rather than simple dichotomous keys, that could benefit citizen scientists.

Generally, the aggregation of results at the NUTS2 level did not show any improvement (Fig. 5), even if one would expect that increased sampling density would improve the results²¹. These results suggested a limitation of the simplified texture-by-feel method used, rather than insufficient data quantity. Overall, our results showed that soil texture is not suitable for direct implementation in soil monitoring and that the ECHOREPO texture classification may be useful for binary analyses that distinguish sandy soils from non-sandy soils, but not when using the current ECHOREPO texture classes as defined.

The moderate agreement between citizen scientists' pH measurements and laboratory-based measurements (Fig. 1A) and LUCAS-derived raster pH values (Fig.3), suggest that the field method combined with the systematic underestimation of 0.4-0.6 pH units represent the main sources of error, in line with the known lower accuracy (± 0.5 pH units) inherent to colourimetric pH measurement methods²². The systematic underestimation was evident for alkaline soils (pH values above 8), suggesting a methodological bias with high pH values having less distinguishable colour transitions, rather than poor skills of citizen scientists to perform the field protocol. The horizontal banding patterns of CS measurements reflect the artefact of the discrete resolution of colorimetric pH strips. This can have affect the RMSE and reduced the $\text{adj}R^2$ when comparing citizen scientists pH measurements to LUCAS-derived raster data, reflecting an inherent constraint of the field method itself rather than a citizen scientist poor execution of the field method. Furthermore, not all of the disagreement between CS pH measurements and LUCAS-derived raster data can be attributed solely to CS error, as the raster dataset itself has an associated uncertainty²³.

The spatial aggregation at the NUTS2 level improved the agreement and reduced part of the background noise associated with individual CS measurements (Fig. 4B). The agreement varied across regions, likely reflecting both the heterogeneity of soils across Europe and differences in how well the citizen science samples represented this heterogeneity at the NUTS2 level. Other factors of variation can be the prior knowledge of citizen scientists, and their training to use the field protocol²¹. Nevertheless, the aggregation at NUTS2 level provided meaningful information at regional level and might be relevant where large number of measurements can compensate part of the deviation. However, in our study the relationship between RMSE on pH and sampling density (Fig.

S3) indicated that aggregating more CS measurements within a region did not consistently reduce the deviation from median LUCAS-derived pH values. Increasing the sampling density is not sufficient to improve data quality, as a systematic negative bias was also observed at NUTS2 level. Overall, the higher agreement observed at the NUTS2 level than at the individual measurement level suggests that citizen scientists' pH measurements are more suitable for broad-scale monitoring than for point-level observations. Following appropriate bias correction, these data could complement existing soil monitoring datasets and support quantitative analyses.

The marked discrepancy between ECHOREPO sites and their nearest LUCAS points reflected differences in land cover distribution (Fig. 2). While ECHOREPO sites were predominantly located in urban areas (31.1%), the nearest LUCAS points were mainly situated on arable land (37.7%) highlighting the differences in the sampling design between the two datasets. LUCAS nearest points, extracted from LUCAS Soil followed a regular grid that covers the European territory selecting the sampling sites based on a random sampling design stratified by land use and topography⁸. In contrast, ECHOREPO site selection is driven by the citizen scientists' location, site accessibility, and personal or local preferences for sampling specific sites. As a result, the distribution of land cover types in ECHOREPO does not reproduce the land cover distribution observed in the LUCAS survey.

The low correspondence between ECHOREPO sites land cover types at the nearest LUCAS point for urban areas (9.6%) further limits the use of LUCAS-derived raster data as reference for ECHOREPO measurements (Fig. S2). This might explain partly the deviation observed between citizen scientists' measurements and LUCAS-derived raster values. Soil properties can vary with land use, management, disturbance and local environmental conditions. Consequently, discrepancies between citizen scientists' measurements and LUCAS-derived raster data may arise not only from errors associated with the citizen science methods (e.g. for pH and soil texture), but also from differences in the types of sites represented in the ECHOREPO and LUCAS datasets.

The high proportion of urban areas in ECHOREPO reflects a bias relative to LUCAS Soil, in which urban areas remain largely underrepresented. Rather than a limitation, however, this can be seen as a valuable complementary contribution. Indeed, urban soil health research remains comparatively limited²⁴, despite the recognised role of urban ecosystems in providing essential ecosystem services, including the provision of clean water and food, control of air pollution, recreational opportunities and climate adaptation²⁵. Our findings indicate that ECHOREPO provides complementary coverage to existing European professional soil surveys by including many urban sites that are

otherwise underrepresented. This complementary dataset could help fill monitoring gaps and guide future research on urban soil challenges.

Lastly, our results on digital soil mapping showed that EU-wide CS measurements could enhance an existing EU soil survey, although the magnitude of this improvement depends on sample size. The enhancement of existing surveys by CS measurements has been shown prominently in the field of ecology^{26,27}, and to a lesser extent in soil science for regional datasets²⁸. Our study extends previous work by quantifying how the contribution of citizen science (CS) soil measurements to digital soil mapping varies regionally across the EU. Our results showed that no single CS implementation consistently improved model performance across all NUTS2 regions, highlighting strong spatial variability in the value of CS data. These regional differences are likely driven by variation in the spatial clustering of CS observations²⁹, differences among surveyors^{30,31}, disparities in feature and geographic coverage relative to the LUCAS survey²⁸ (Fig. 2), and biases in the CS measurements (Fig. 1A).

Regarding the latter, the comparison between CS and ISO pH showed a bias particularly for low pH (Fig. 1A). Consistent with this, NUTS2 regions with significant WSRT statistics had higher median pH values (Fig. S4a), suggesting that the lower bias at higher pH may have contributed to the improved model performance. The PI90 weighting may also have reduced the influence of biased measurements, as the CS observations selected most frequently had higher pH values than the full set of CS observations (Fig. S4c–d). Finally, the qRF model may have further reduced the effect of measurement bias through spatial smoothing, especially given the high density of CS observations relative to the resolution of the environmental covariates. For example, 36% of raster cells containing CS data included more than one CS measurement.

By adding CS measurements to the LUCAS data, we identified a sample size- dependent bias-variance trade-off. This trade-off was reduced by selecting CS measurements using metrics, supporting our hypothesis that prediction uncertainty of the LUCAS model provides useful information for selecting CS observations. These results are consistent with previous studies, which found that prior information can be used to guide the selection of CS data^{28,30}. However, the effect was limited up to a sample size of ($N=2k$), after which the positive effect of the CS measurements decreased (R^2). Notably, the inclusion of CS data improved the 90th prediction interval statistics (IS, PICP) suggesting that citizen science measurements contribute not only to prediction accuracy but also to more reliable uncertainty quantification. Lastly, not all weighting metrics for CS data led to improved model performance. In our study, the lower performance of the PIQ50

weighting metric may be explained by its greater variability across iterations compared to other metrics (Fig. S5).

Our results show that the value of citizen science for European soil monitoring depends on both the soil property and how the data are integrated with existing monitoring networks. CS pH measurements captured broad spatial patterns, particularly when aggregated at regional scales or calibrated against reference datasets. In contrast, SOC estimates were more suitable as qualitative indicators, while soil texture was only informative when simplified into broader classes, such as sandy versus non-sandy soils, before integration with European soil monitoring datasets. Despite these limitations, ECHOREPO fills an important gap by increasing the spatial density of soil observations and including sites that are underrepresented in conventional monitoring, particularly urban areas. This is important because urban soils are highly heterogeneous but remain poorly represented in systematic soil surveys.

Our digital soil mapping analysis further showed that CS pH measurements can improve predictions when integrated with the LUCAS Soil survey. The reduction in prediction error and the improvement in prediction intervals, however, were only observed when CS observations were selected using weighting metrics derived from the baseline LUCAS model. Moreover, the effect of adding CS data differed across the EU and depended on sample size, reflecting a bias-variance trade-off that needs to be considered when integrating citizen science data into digital soil mapping.

Overall, our findings suggest that ECHOREPO should be considered a complementary source of soil information rather than a replacement for established monitoring programmes such as LUCAS Soil. Rather than increasing sampling density indiscriminately, the greatest benefit comes from integrating citizen science observations where they add new information and complement existing monitoring. This provides a practical basis for incorporating citizen science into future European soil monitoring while extending coverage to environments that are currently underrepresented.

Materials and Methods

Soil sampling

A discrete soil sampling strategy was applied. Citizen scientists excavated a soil pit of 30 × 30 × 30 cm using a metal hand trowel. The excavated soil was subsequently homogenised thoroughly prior to the determination of soil pH, organic carbon content, and soil texture. The samples were collected by citizen scientists between 25/05/2025 and 22/06/2026.

A set of 100 samples was collected in Italy by citizens scientists at the same sampling points and at the same time as the main soil samples. The same sampling protocol was applied. Approximately 0.5 kg of homogenised soil was collected from each sampling point and sent to the laboratory for subsequent analysis of pH, organic carbon content and soil texture using laboratory-based methods.

Citizen science methods

Soil pH. Citizen scientist measured soil pH using pH test strip (MQuant®, Merck, Germany), and a standard soil-to-water ratio of approximately 1:2.5 (w:v). Approximately a spoonful of homogenised soil (approx. 4 g) was added to a centrifuge tube containing 10 mL of distilled water, until the mixture reached 14 mL. The tube was shaken for 30 seconds, then placed upright to allow the mixture to settle for 5 minutes. The coloured end of the pH paper strip was dipped into the solution for 30 s, then the colour of the strip was compared to the pH colour chart (Fig. S6).

Soil organic carbon content. Soil organic carbon content was estimated by citizen scientist using a soil colour comparison chart, assigning values ranging from 1.5 to 5% (Fig S7). Values exceeding 5% were recorded as “Other”, corresponding to organically rich soils characterised by high organic matter content and very dark colour.

Soil texture. Soil texture assessment was assessed by conducting the “texture-by-feel” method, adapted from USDA soil quality guideline³² simplifying the texture classes up to seven (i.e. Sandy, silty, clayey, sandy loam, sandy clay, and silty clay; Fig S8). For organic rich soils, texture was recorded as “Other” when it could not be reliably classified due to limited presence of clay, sand and/or silt particles.

Laboratory-based methods

Soil pH. Soil pH was determined using a portable pH meter (pH 7 Vio, Metrocal, Italy) after extraction using a soil-to-solution ratio of 1:2.5. Briefly, 10 g of oven-dried, sieved soil (<2 mm) were mixed with 25 mL of deionised water, the suspension was shaken for 60 min and then centrifuged at 2000 g for 3 min. pH-H₂O was subsequently determined directly in the supernatant.

Soil organic carbon. Oven-dried soil was ball-milled for 6 min at 30 Hz (Mixer Mill MM 400, Retsch, Italy), and approximately 0.2 g of homogenised material was weighed into silver capsules. Samples were acidified with 1 mol L⁻¹ HCl and heated to remove carbonates before combustion in an elemental analyser (EA Flash 1112, Thermo Scientific, Germany) for TOC determination. The oxidation and reduction furnaces operated at 1020 °C and 900 °C, respectively, and H₂O was removed with a Mg(ClO₄)₂ trap.

Soil texture. Soil texture was determined using a combined sieve–hydrometer method after dispersion with sodium pyrophosphate ($\text{Na}_4\text{P}_2\text{O}_7$). Briefly, 100 g of oven-dried, sieved soil were placed in a 1 L graduated cylinder with 100 mL of distilled water and 25 mL of $0.1 \text{ mol L}^{-1} \text{ Na}_4\text{P}_2\text{O}_7$, shaken thoroughly, and brought to volume. After a second shake, hydrometer readings were taken at 30 s (sand and silt fractions) and 2 h (clay fraction), together with suspension temperature. The proportions of sand, silt and clay were then calculated using standard particle-size equations and then the texture was assigned using the USDA triangle.

Data processing and analysis

ECHOREPO. ECHOREPO³³ was developed as the digital infrastructure supporting the collection, curation, access and reuse of citizen science soil observations generated within the project. The platform is an open-source and reusable codebase grounded in established web, database, authentication and data-publication technologies. ECHOREPO is continuously updated as new samples are registered when citizen scientists complete the sampling and data collection. The ECHOREPO dataset comprises soil data collected in the field by citizen scientists, at the sampling location of their choice. The dataset includes information on soil pH, soil organic carbon, soil texture, soil structure, soil biodiversity (expressed as earthworms), soil contamination (expressed as number of metal and plastic debris), GPS coordinates of the sampling site and photographs of the soil sample, landscape and vegetation cover at the sampling site. Additional information through comments can be found in the dataset and were added by the citizen scientists when collecting their samples. The dataset was accessed on 22/06/2026 and contained 6954 unique entries. Data were screened before analysis, and entries were discarded when (i) the GPS coordinates contained errors, (ii) the sample collection was incomplete, (iii) the samples corresponded to pot experiments rather than field soil samples. Overall, 225 data were excluded, resulting in 6729 entries for further analysis. The ECHOREPO soil sampling locations, number of samples per country and number of samples per major land covers are shown in Fig S9, Tables S1 and S2.

Comparison of ECHOREPO citizen science measurements with laboratory-based measurements. The deviation between citizen scientist measured soil parameters (pH and organic carbon content) and the laboratory-based measurements was determined by calculating the regression slope and coefficient as well as the root mean square error (RMSE), the relative RMSE, the mean bias error (MBE) and the relative MBE. The agreement between soil texture classes assessed by citizen scientist and laboratory-based method was assessed by using the Weighted Kappa coefficient. Kappa coefficient can range from -1 to 1, with values < 0 showing poor agreement, values between 0 and 0.2 indicate slight

agreement, values between 0.2 and 0.4 indicate fair agreement, values between 0.4 and 0.6 indicate moderate agreement, values between 0.6 and 0.8 indicate substantial agreement and values between 0.8 and 1 indicate almost perfect agreement³⁴.

Comparison of ECHOREPO citizen science measurements with LUCAS 2018 data. The LUCAS 2018 dataset used in this study consists of records from the 2018 sampling campaign, based on a random sampling design stratified by land use and topography. Soil samples were taken to be representative of 0–20 cm depth and an areal extent of 16 m², see Orgiazzi et al. (2018)⁸ for further details. The bulked soil samples were air-dried and sieved to their < 2 mm fractions. Soil analytical data for clay, organic carbon and pH in H₂O was then determined following ISO protocols.

For each ECHOREPO and nearest LUCAS sampling point, the land cover types were extracted from CORINE (Coordination of Information on the Environment) Land Cover 2018³⁵ to assess the composition in land cover from the two datasets. The second level of CLC was used. The classes 11,12,13 and 14 were grouped all together in unique classes renamed “Urban areas” as in the ECHOREPO dataset, class 11 (Urban fabric class in CLC) represented 25% of land cover and 12, 13 and 14 represented 3, 0.4 and 2% respectively.

The next steps of our assessment consisted of comparing the ECHOREPO data with the LUCAS 2018 data. We implemented this in three distinct steps. First, for each ECHOREPO sampling point, the nearest LUCAS sampling point was identified. Among the available LUCAS 2018 soil properties, only pH measured in water, pH-H₂O, was retained for further analysis. Secondly, to map the agreement on pH-H₂O and soil texture between ECHOREPO and LUCAS datasets, we used the derived-raster data from LUCAS dataset^{23,36}. For each ECHOREPO sampling point, raster pH values and soil texture classes were extracted using a buffer of 1000 m radius. Since ECHOREPO use a simplified seven soil texture classification, LUCAS-derived raster texture data were harmonised accordingly before comparison (Table S3) and the “Other” class from ECHOREPO was considered as non-available data.

For the spatial distribution of pH RMSE and Kappa coefficient for soil texture, we used the Nomenclature of Territorial Units for Statistics (NUTS) which is a system developed by Eurostat to divide the European Union’s territory into different regions for the purpose of statistics. The system has three levels NUTS1 (large regions with population of 3-7 million inhabitants), NUTS2 (group of counties and unitary authorities with a population of 0.8-3 million inhabitants), and NUTS3 regions (counties of 150-800 thousand inhabitants)³⁷. The NUTS2 regions were used for spatial representation of data (i) to ensure adequate sample density per region for robust statistical estimation and (ii) to be consistent with precedent digital soil mapping using LUCAS Soil data³⁸. We assessed the relationship between pH RMSE with sampling density at NUTS2 regions level. Sampling density was calculated as

the number of samples per NUTS2 region divided by the region's area, expressed per 100 km². For the spatial distribution of pH RMSE and Kappa coefficient for soil texture at NUTS 2 level, no sample size threshold was applied. The NUTS2 regions with fewer than 5 paired samples represented 26.7 % of the entries for RMSE computed for pH and 26.4 % for Kappa coefficient computed for soil texture.

Digital soil mapping to integrate ECHOREPO citizen science measurements with LUCAS. We then investigated to what extent the ECHOREPO data could be integrated with the LUCAS survey when creating a digital soil map of pH. For this purpose, we combined the ECHOREPO citizen-science (CS) pH measurements with the LUCAS 2018 dataset to fit a quantile random-forest (qRF) model.

As a case-study area, we used the NUTS2 regions that contained CS measurements (see Suppl. Material) and the associated LUCAS 2018 data. For both datasets we extracted the same covariate layers that were used in Ballabio et al. (2019) to map pH across the European Union (Table S4) with one additional covariate: the aridity index (potential evapotranspiration / precipitation). We calculated the mean aridity index over the 01-01-2001—01-01-2021 period, as in Breure et al. (2025)³⁹, using the TerraClim dataset available in Google Earth Engine⁴⁰. Given that the covariate layers were at 500-meter resolution, we averaged the CS locations that fell within the same raster cell, giving us the final CS dataset (n = 4933).

The detailed steps were as follows: first, the LUCAS data were split by taking a random sample of 75% for training, where the remaining 25% was the hold-out test set. Second, given the large number of covariates, we applied recursive feature elimination with 10-fold repeated cross-validation to select the optimal covariates that minimized the RMSE. Thirdly, we used the fitted qRF to map the 5th, 50th and 95th quantiles (q5, q50, q95) of pH and extracted these predictions for the CS sites (Fig. 10).

We hypothesized that CS measurements at locations where the LUCAS map was uncertain would contain more useful information. To test this hypothesis, we calculated a range of uncertainty metrics based on the 90th prediction interval (q95-q5) of the CS pH data. These were: prediction-interval scaled by the median (q50), the interval score (IS) (Gneiting and Raftery, 2007), and whether the CS measurement was covered within the prediction interval (see Suppl. Material for further details). These metrics were combined in a full-factorial manner to generate case-weighting schemes for the CS measurements (Fig. S10).

The CS measurements, which had not been used so far, were then re-arranged in decreasing order based on the weighting scheme. We then tested the effect of selecting by weighting schemes, by selecting a sample size $N \in \{1000, 2000, 3000, 4000, 4933\}$ from the CS measurements. We also tested for an unweighted (equal) scenario,

where we obtained a random sample of size N , so we could assess the effect of applying the weighting. The selected samples were added to the LUCAS training data, and we refitted the qRF model, using the same covariates as the original model. The entire procedure described above was repeated 100 times, to minimize differences between models due to a specific feature- and/or geographic distribution in training/testing data.

Once all the results were obtained, we assessed the predictions on the hold-out test set from the various model configurations and iterations. We assessed these using the following mean predictions metrics: the total error (RMSE + |Bias|), coefficient of determination (R^2) and the ratio of performance to inter-quartile range (RPIQ)⁴¹, and the IS and the prediction interval coverage probability (PICP)⁴². We calculated these metrics both for the entire test set, and also by NUTS2 region to quantify the local effect of adding CS measurements. The NUTS2 region comparison was only done for the best performing weight- and sample size combination, compared to the LUCAS data alone. We performed pair-wise testing (Wilcoxon Signed Rank Test, WSRT) on the differences in total error distributions across 100 iterations between the CS and LUCAS implementation. We did this where: i.) the number of test samples in the NUTS2 exceeded two, and ii.) the frequency with which a NUTS2 region appeared in the test set exceeded two, as required by the WSRT (see Suppl. Material for further details).

All statistical analysis were performed using R software⁴³. In addition to base R functions the following packages were used: ranger⁴⁴, terra⁴⁵, sf⁴⁶, irr⁴⁷, geosphere⁴⁸, eurostat⁴⁹, and tidyverse⁵⁰.

Data availability

The ECHOREPO is available in the Zenodo repository at <https://doi.org/10.5281/zenodo.21105834>.

References

1. Panagos, P. *et al.* How the EU Soil Observatory is providing solid science for healthy soils. *European J Soil Science* **75**, e13507 (2024).
2. Directive (EU) 2025/2360 of the European Parliament and of the Council of 12 November 2025 on soil monitoring and resilience (Soil Monitoring Law).

- 616 3. *The State of Soils in Europe: Fully Evidenced, Spatially Organised Assessment of the*
617 *Pressures Driving Soil Degradation*. (Publications Office, Luxembourg, 2024).
618 doi:10.2760/5897030.
- 619 4. Morvan, X. *et al.* Soil monitoring in Europe: A review of existing systems and
620 requirements for harmonisation. *Science of The Total Environment* **391**, 1–12 (2008).
- 621 5. Van Leeuwen, J. P. *et al.* Gap assessment in current soil monitoring networks across
622 Europe for measuring soil functions. *Environ. Res. Lett.* **12**, 124007 (2017).
- 623 6. Meurer, K. H. E. *et al.* How does national SOC monitoring on agricultural soils align
624 with the EU strategies? An example using five case studies. *European Journal of Soil*
625 *Science* **75**, e13477 (2024).
- 626 7. Froger, C. *et al.* Comparing LUCAS Soil and national systems: Towards a harmonized
627 European Soil monitoring network. *Geoderma* **449**, 117027 (2024).
- 628 8. Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A. & Fernández-Ugalde, O. LUCAS Soil,
629 the largest expandable soil dataset for Europe: a review. *European J Soil Science* **69**,
630 140–153 (2018).
- 631 9. Sapkota, M., Young, J., Slaughter, L., Acosta-Martínez, V. & Coldren, C. Soil microbial
632 biomass and composition from urban landscapes in a semiarid climate. *Applied Soil*
633 *Ecology* **158**, 103810 (2021).
- 634 10. Mason, E. *et al.* Participatory soil citizen science: An unexploited resource for
635 European soil research. *European Journal of Soil Science* **75**, e13470 (2024).
- 636 11. Charvát, K. *et al.* Citizen Science for Soil Monitoring and Protection in Europe:
637 Insights from the PREPSOIL Project Under the European Soil Mission. *Sustainability*
638 **17**, 5042 (2025).

- 639 12. Echo Horizon. *EchoSoil* <https://echosoil.eu/>.
- 640 13. Ferrando Jorge, N., Clark, J., Cárdenas, M. L., Geoghegan, H. & Shannon, V.
641 Measuring Soil Colour to Estimate Soil Organic Carbon Using a Large-Scale Citizen
642 Science-Based Approach. *Sustainability* **13**, 11029 (2021).
- 643 14. Aksoy, E., Yigini, Y. & Montanarella, L. Combining Soil Databases for Topsoil
644 Organic Carbon Mapping in Europe. *PLoS ONE* **11**, e0152098 (2016).
- 645 15. Torrent, J. & Barrón, V. Laboratory Measurement of Soil Color: Theory and
646 Practice. in *SSSA Special Publications* (eds Bigham, J. M. & Ciolkosz, E. J.) vol. 31 21–
647 33 (Wiley, 1993).
- 648 16. Schwertmann, U. Relations Between Iron Oxides, Soil Color, and Soil Formation.
649 in *SSSA Special Publications* (eds Bigham, J. M. & Ciolkosz, E. J.) vol. 31 51–69 (Wiley,
650 1993).
- 651 17. SÁnchez-Marañón, M., Soriano, M., Melgosa, M., Delgado, G. & Delgado, R.
652 Quantifying the effects of aggregation, particle size and components on the colour of
653 Mediterranean soils. *European J Soil Science* **55**, 551–565 (2004).
- 654 18. echosoil/soil-color-orgc., <https://github.com/echosoil/soil-color-orgc>. echosoil
655 (2026).
- 656 19. Richer-de-Forges, A. C. *et al.* Hand-feel soil texture and particle-size distribution
657 in central France. Relationships and implications. *CATENA* **213**, 106155 (2022).
- 658 20. Salley, S. W. *et al.* A Comparison of Soil Texture-by-Feel Estimates: Implications
659 for the Citizen Soil Scientist. *Soil Science Soc of Amer J* **82**, 1526–1537 (2018).
- 660 21. Kosmala, M., Wiggins, A., Swanson, A. & Simmons, B. Assessing data quality in
661 citizen science. *Frontiers in Ecol & Environ* **14**, 551–560 (2016).

- 662 22. Mosley, L. M., Rengasamy, P. & Fitzpatrick, R. Soil pH : Techniques, challenges
663 and insights from a global dataset. *European J Soil Science* **75**, e70021 (2024).
- 664 23. Ballabio, C. et al. Mapping LUCAS topsoil chemical properties at European scale
665 using Gaussian process regression. *Geoderma* **355**, 113912 (2019).
- 666 24. Vuono, M., Di Lonardo, D. P., Schwartz, C. & Terribile, F. Advancing urban soil
667 health: challenges, knowledge gaps, and future research perspectives. *Landscape
668 and Urban Planning* **266**, 105534 (2026).
- 669 25. Gómez-Baggethun, E. et al. Urban Ecosystem Services. in *Urbanization,
670 Biodiversity and Ecosystem Services: Challenges and Opportunities: A Global
671 Assessment* (eds Elmqvist, T. et al.) 175–251 (Springer Netherlands, Dordrecht, 2013).
672 doi:10.1007/978-94-007-7088-1_11.
- 673 26. Pocock, M. J. O., Tweddle, J. C., Savage, J., Robinson, L. D. & Roy, H. E. The
674 diversity and evolution of ecological and environmental citizen science. *PLoS ONE* **12**,
675 e0172579 (2017).
- 676 27. Lowe, S. K., Mason, B. M., Morales, N. A. & Callaghan, C. T. Participatory citizen
677 science data complements agency-collected data for species inventories. *Ecol Sol
678 and Evidence* **7**, e70173 (2026).
- 679 28. Richer-de-Forges, A. C. et al. Does digital soil mapping prediction performance of
680 soil texture improve when adding uncertain field texture estimates? A study based on
681 clay content. *Geoderma* **456**, 117277 (2025).
- 682 29. Isaac, N. J. B. & Pocock, M. J. O. Bias and information in biological records. *Biol J
683 Linn Soc* **115**, 522–531 (2015).

- 684 30. Bird, T. J. *et al.* Statistical solutions for error and bias in global citizen science
685 datasets. *Biological Conservation* **173**, 144–154 (2014).
- 686 31. Eymard, A. *et al.* Exploring the untapped potential of hand-feel soil texture data
687 for enhancing digital soil mapping: Revealing hidden spatial patterns from field
688 observations. *Geoderma* **441**, 116769 (2024).
- 689 32. NRCS Publications - Item Detail.
690 <https://nrcspad.sc.egov.usda.gov/DistributionCenter/product.aspx?ProductID=385>.
- 691 33. osychenko. echosoil/echorepo-lite: ECHOREPO v0.2.0 – public repository, map
692 visualisation and data access layer. <https://doi.org/10.5281/zenodo.21105834> (2026)
693 doi:10.5281/zenodo.21105834.
- 694 34. Landis, J. R. & Koch, G. G. The Measurement of Observer Agreement for
695 Categorical Data. *Biometrics* **33**, 159 (1977).
- 696 35. EEA geospatial data catalogue.
697 [https://sdi.eea.europa.eu/catalogue/srv/eng/catalog.search#/metadata/5a5f43ca-](https://sdi.eea.europa.eu/catalogue/srv/eng/catalog.search#/metadata/5a5f43ca-1447-4ed0-b0a6-4bd2e17e4f4d)
698 [1447-4ed0-b0a6-4bd2e17e4f4d](https://sdi.eea.europa.eu/catalogue/srv/eng/catalog.search#/metadata/5a5f43ca-1447-4ed0-b0a6-4bd2e17e4f4d).
- 699 36. Ballabio, C., Panagos, P. & Monatanarella, L. Mapping topsoil physical properties
700 at European scale using the LUCAS database. *Geoderma* **261**, 110–123 (2016).
- 701 37. Becker, S. O., Egger, P. H. & von Ehrlich, M. Going NUTS: The effect of EU
702 Structural Funds on regional performance. *Journal of Public Economics* **94**, 578–590
703 (2010).
- 704 38. Panagos, P., Ballabio, C., Yigini, Y. & Dunbar, M. B. Estimating the soil organic
705 carbon content for European NUTS2 regions based on LUCAS data collection.
706 *Science of The Total Environment* **442**, 235–246 (2013).

- 707 39. Breure, T. S. *et al.* Revisiting the soil carbon saturation concept to inform a risk
708 index in European agricultural soils. *Nat Commun* **16**, 2538 (2025).
- 709 40. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a
710 high-resolution global dataset of monthly climate and climatic water balance from
711 1958–2015. *Sci Data* **5**, 170191 (2018).
- 712 41. Bellon Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J. M. & Mcbratney,
713 A. Prediction of soil attributes by NIR spectroscopy. A critical review of chemometric
714 indicators commonly used for assessing the quality of the prediction. *Trends in*
715 *Analytical Chemistry* **29**, 1073–1081 (2010).
- 716 42. Zhang, L. & Garcia-Munoz, S. A comparison of different methods to estimate
717 prediction uncertainty using Partial Least Squares (PLS): A practitioner’s perspective.
718 *Chemometrics and Intelligent Laboratory Systems* **97**, 152–158 (2009).
- 719 43. R: The R Project for Statistical Computing. <https://www.r-project.org/>.
- 720 44. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for
721 High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**, 1–17 (2017).
- 722 45. Hijmans, R. J., Brown, A. & Barbosa, M. terra: Spatial Data Analysis. 1.9-34
723 <https://doi.org/10.32614/CRAN.package.terra> (2020).
- 724 46. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data.
725 *The R Journal* **10**, 439–446 (2018).
- 726 47. Gamer, M., Lemon, J., Fellows, I. & Singh, P. irr: Various Coefficients of Interrater
727 Reliability and Agreement. (2026).
- 728 48. Hijmans, R. J., Karney (GeographicLib), C., Williams, E. & Vennes, C. geosphere:
729 Spherical Trigonometry. (2026).

49. Lahti, L. *et al.* eurostat: Tools for Eurostat Open Data. (2023).
50. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

Acknowledgments

The authors express their sincere gratitude to the citizens who have engaged throughout the ECHO process.

Funding statement

C. Laurent, C. Cappello, T. Mimmo, and O. Osychenko received funding from the Horizon Europe Mission Soil project ECHO, *Engaging Citizens in Soil Science: The Road to Healthier Soils*, a Research and Innovation Action co-funded by the European Union under the Horizon Europe programme (GA. NO. 101112869), and by UK Research and Innovation (UKRI) (GA NO. 10068004). The LUCAS Soil sample collection is supported by EUROSTAT and the following Directorate-Generals of the European Commission: Environment (DG-ENV), Agriculture and Rural Development (DG-AGRI) and Climate Action (DG-CLIMA). The specific version of the ECHOREPO citizen science data used in this study is available from the European Soil Data Centre (ESDAC): [hyperlink to be inserted].

Authors contribution

Céline Laurent: conceptualization, data collection and curation, methodology, software, visualization, writing - original draft. Claudia Cappello: conceptualization, data collection and curation, methodology, software, visualization, writing - original draft, project administration. Timo Breure: conceptualization, data collection and curation, methodology, software, visualization, writing - original draft. Oleg Osychenko: data curation, writing – original draft. Panos Panagos: conceptualization, methodology, writing - revision. Cristiano Ballabio: conceptualization, methodology, data collection and curation. Arwyn Jones: conceptualization, methodology, writing - revision. Tanja Mimmo: conceptualization, methodology, software, visualization, writing - original draft, supervision, validation, project administration, resources.

763 **Competing interest**

764 The authors declare no competing interests.

765 **Material and correspondence**

766 All requests should be addressed to Céline Laurent (celine.laurent@unibz.it)

767

768

769

770

771

772

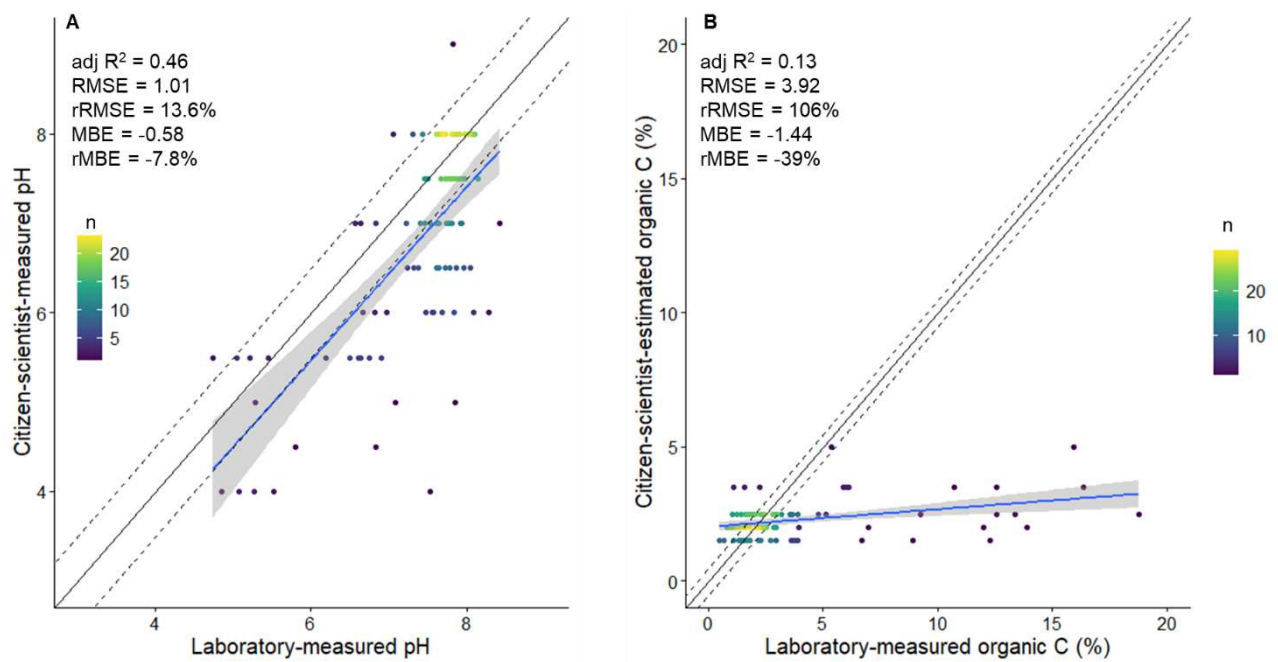


Figure 1: Comparison of citizen scientist measurements with laboratory-based measurements. pH values (A) and soil organic carbon content (B) measured using citizen science methods as a function of the same parameters measured using laboratory-based methods. For each data point, the colour represents the number of samples (n), with the darker colour the fewer samples for each single value. The blue line, the thin solid line and the dashed lines represent the regression line with 95% confidence interval (grey shaded area), the 1:1 line and the ± 0.5 pH unit or organic C (%), respectively. The adjusted coefficient of the regression (adj R^2), the root mean square error (RMSE), the relative root mean square error (rRMSE), the mean bias error (MBE), and the relative mean bias error (rMBE) are calculated.

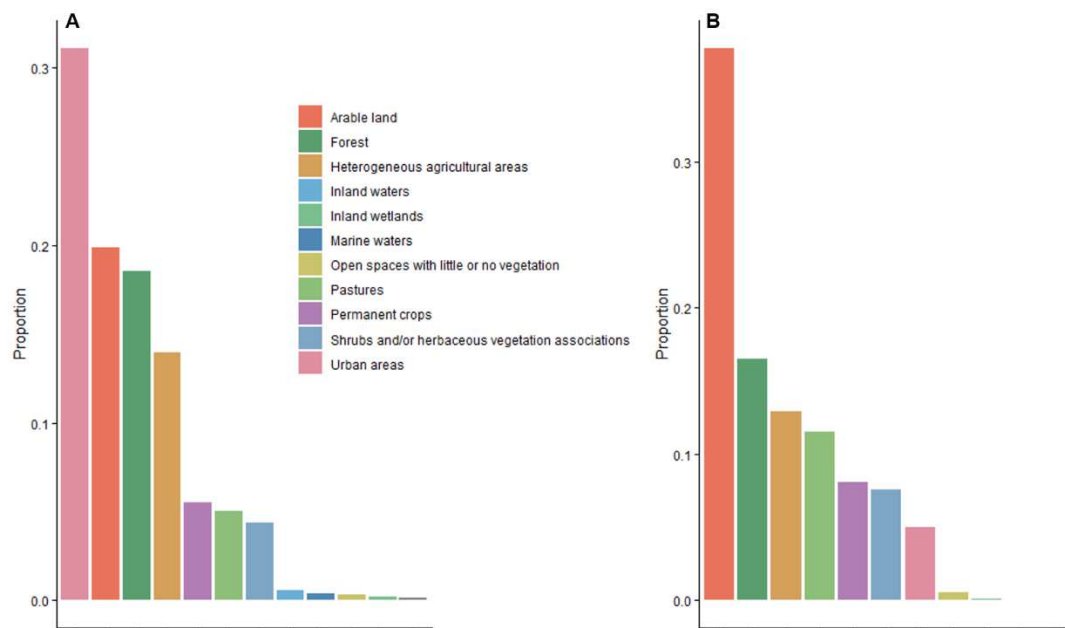


Figure 2: Proportion of land cover types in ECHOREPO dataset (A) and LUCAS 2018 dataset (B). Each colour represents a different land cover type extracted from CORINE Land Cover 2018 raster (100 m resolution).

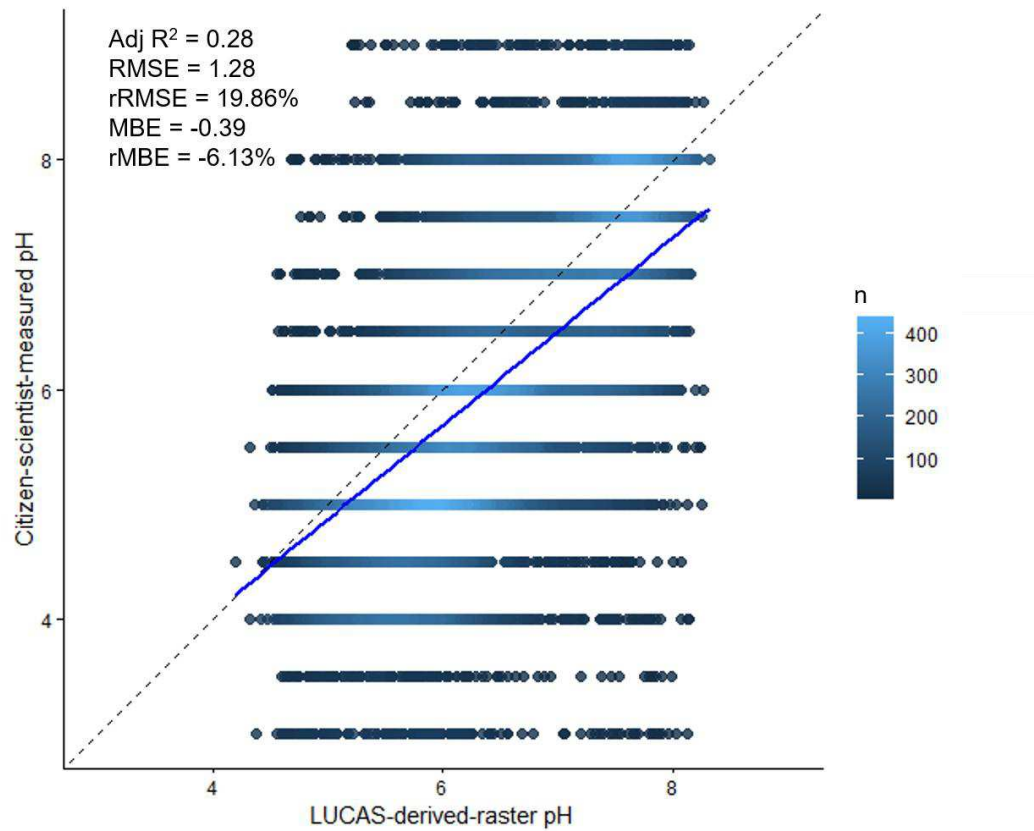
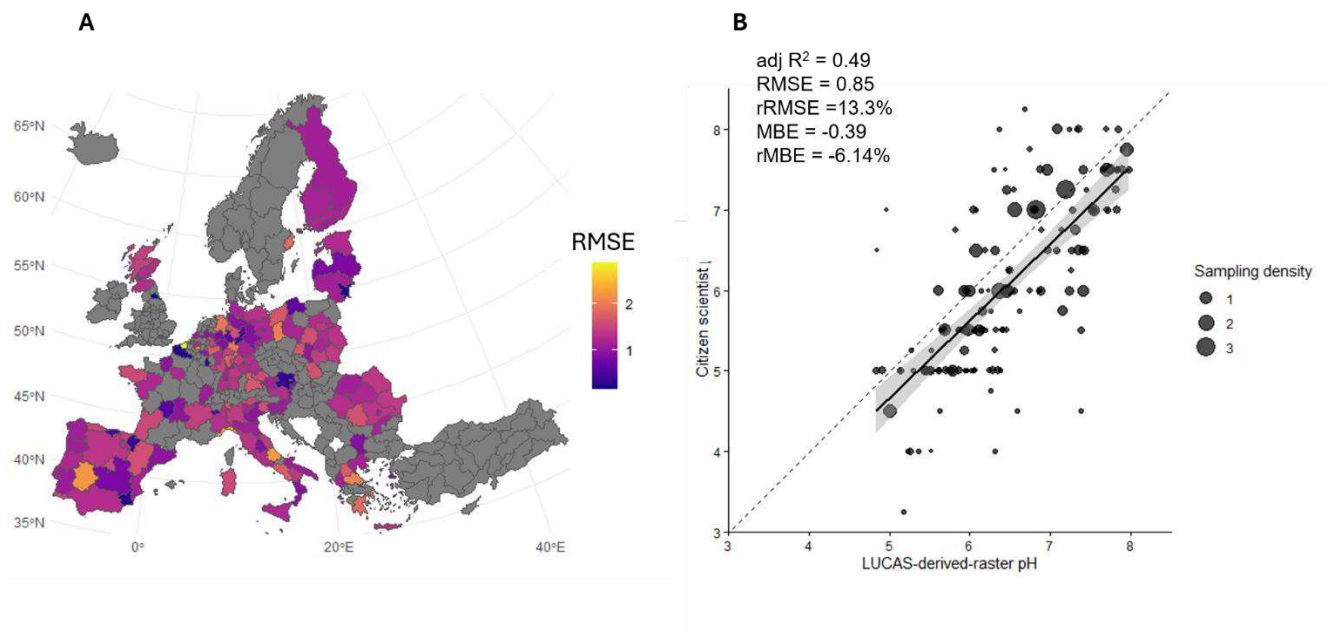


Figure 3: pH values measured by ECHO citizen science method as a function of LUCAS-derived raster pH values. For each data point, the colour represents the number of samples (n), with the darker colour the fewer samples for each single value. The black dashed line and the blue thin line represent the 1:1 line and the regression line, respectively.

36



37

38 **Figure 4:** Aggregation of pH measurements and root mean square error (RMSE) at NUTS2
 39 level. Spatial distribution (A) of the calculated RMSE between citizen scientists soil pH
 40 measurements and LUCAS-derived-raster pH data at NUTS2 level across Europe and
 41 scatter plot of median pH values measured by citizens scientists as a function of LUCAS-
 42 derived raster median pH values (B). On the map, the darker the colour, the lower the
 43 RMSE. Grey zone corresponds to areas where no ECHO samples were collected when the
 44 data were extracted from ECHOREPO. For the scatter plot, each point represents a NUTS2
 45 region ($n = 146$), with point size representing the sampling density (number of samples per
 46 100 km^2) of a region. The black thin line represents the regression line with 95%
 47 confidence interval (grey shaded area).

48

49

50

51

52

53

54

55

56

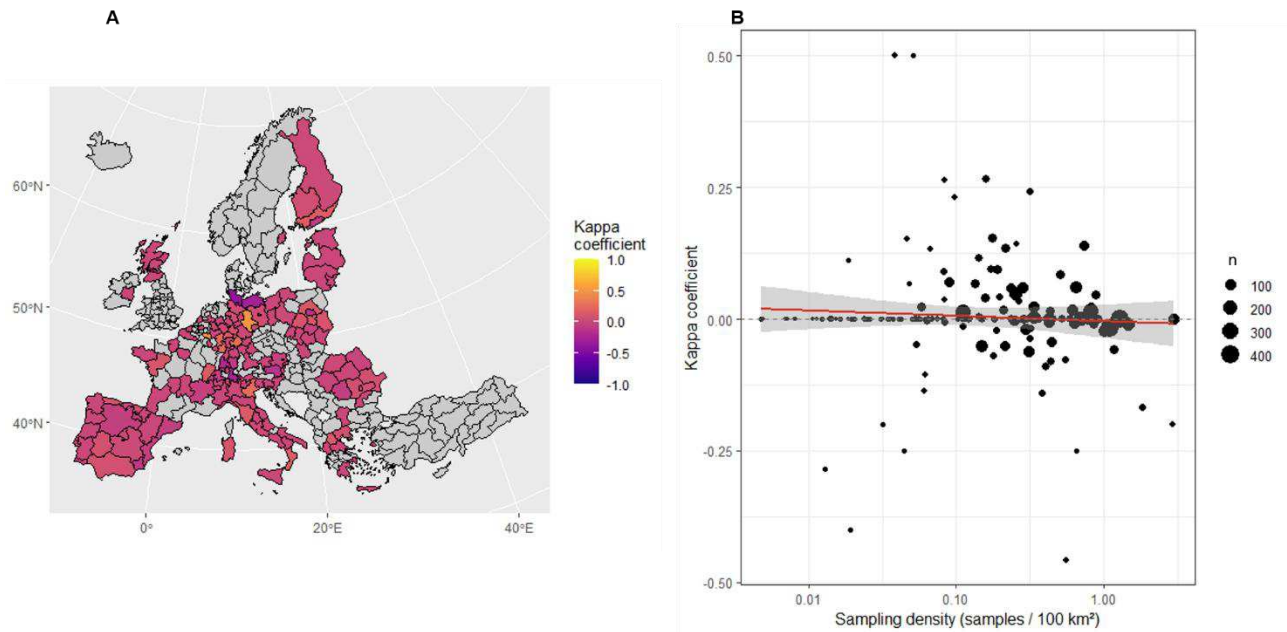


Figure 5: Aggregation of Kappa coefficient at NUTS2 level. Spatial distribution (A) of the calculated Kappa coefficient between citizen scientists soil texture assessment and LUCAS-derived-raster texture data at NUTS2 level across Europe and scatter plot of Kappa coefficient at NUTS2 level as a function of the sampling density (number of samples per 100 km²) (B). On the map, the darker the colour, the lower the Kappa coefficient. Grey zone corresponds to areas where no ECHO samples were collected when the data were extracted from ECHOREPO. For the scatter plot, each point represents a NUTS2 region (n = 148), with point size representing the sampling density (number of samples per 100 km²) of a region. The red thin line represents the regression line with 95% confidence interval (grey shaded area).

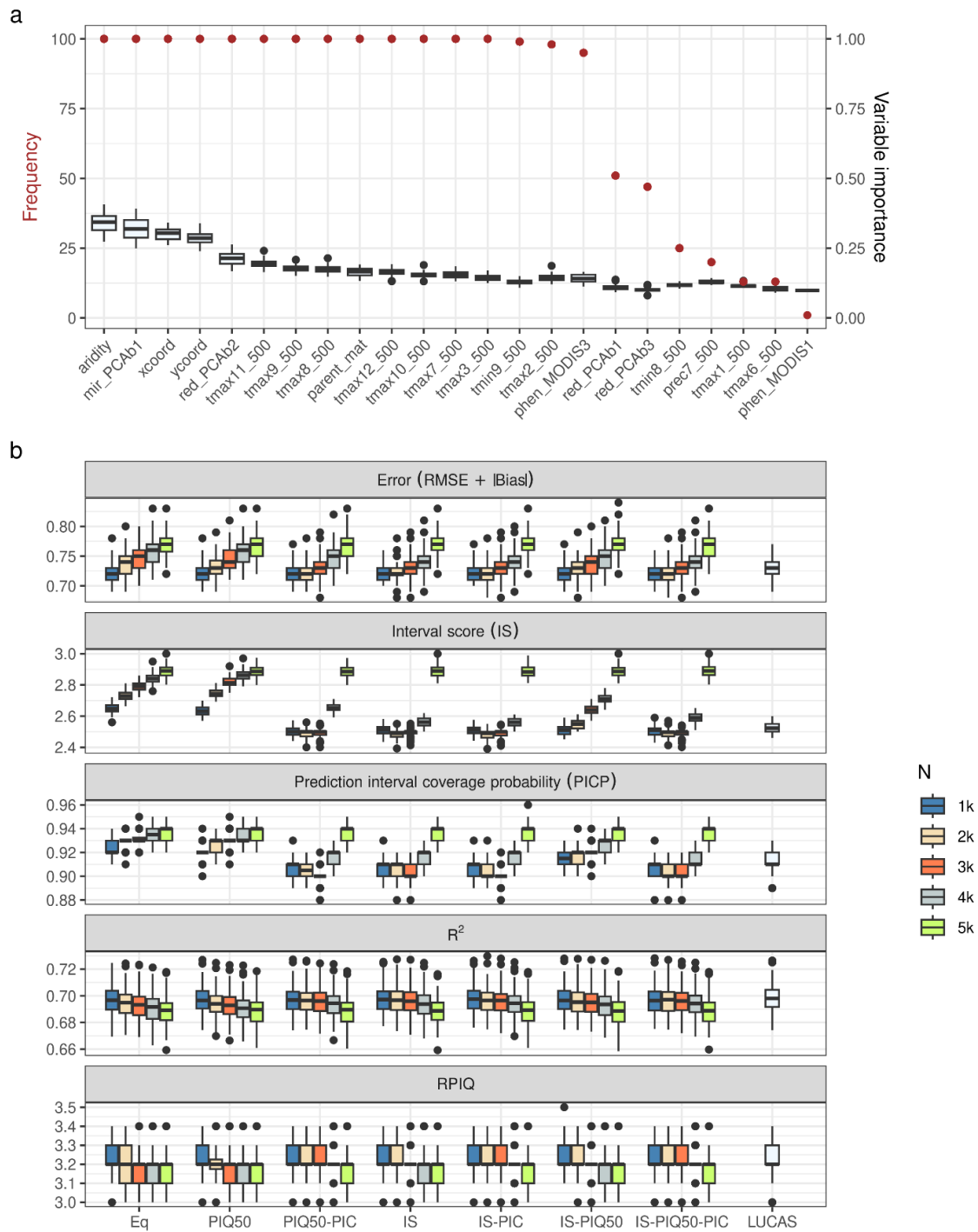
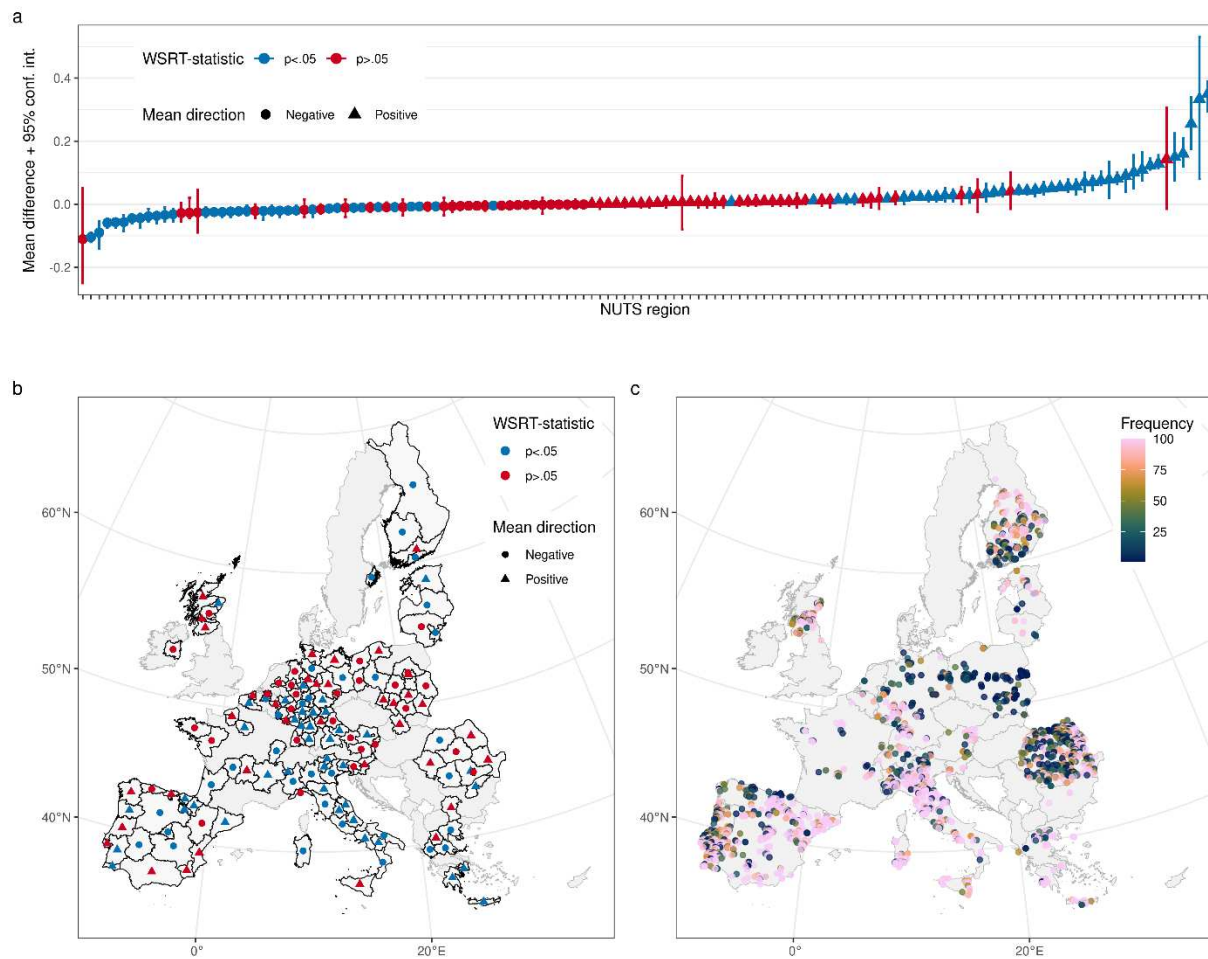


Figure 6. A. The distribution of variable importance for the covariates selected for the quantile regression forest models. The distributions represent the variation across all 100 iterations, the brown dots represent the frequency with which a covariate was selected. **B.** the distribution of prediction metrics on the test set across all 100 iterations, for varying sample sizes (N). LUCAS represents the model without any citizen science (CS) data, Eq. is with CS data added through a random sample, the remaining categories are different weighting schemes used in the selection of CS data (see Methods).

85
86



87

88 **Figure 7. A.** the Wilcoxon signed-rank test (WSRT) results at NUTS2 level. Symbols represent
89 the direction of the effect (positive- or negative) and colors whether the test was significant at
90 an $\alpha = 0.05$ level. Lines correspond to the 95% confidence intervals. **B.** spatial representation of
91 the WSRT results at NUTS2 level, where symbols and colors are follow the same structure as
92 panel a. **C.** the frequency in which a citizen-science measurement has been selected across the
93 100 iterations. The citizen-science data in all panels is represented by the IS-PIQ50-PIC
94 weighting implementation, considering a sample size ($N=2000$).

95

1 **Table 1:** Confusion matrix for soil texture assessed by citizen scientists and soil texture
 2 based on laboratory method. Rows correspond to soil texture classes from citizen
 3 scientists' observations and columns correspond to soil texture classes from laboratory
 4 measurements. Diagonal values correspond to correct match of texture classes and off-
 5 diagonal values indicate disagreement among texture classes. Values indicate the
 6 number of samples assigned to each texture class combination. Laboratory-based
 7 texture classes were harmonized to fit the ECHO seven texture classes.

	Clay loam	Clay	Sand	Sandy clay	Sandy loam	Silty	Silty clay	Other
Clay loam	1				5			
Clay	3	0			6			
Sand			2		7			
Sandy clay	3			0	10			
Sandy loam	1				19			
Silty	2		1	1	15	0		
Silty clay	5				17		0	
Other			1		1			0

8

9

10

11

12

13

14

15

Table 2: Confusion matrix for soil texture assessed by citizen scientists and LUCAS-derived raster texture. Rows correspond to soil texture classes from citizen scientists' observations and columns correspond to LUCAS-derived raster texture. Diagonal values correspond to correct match of texture classes and off-diagonal values indicate disagreement among texture classes. Values indicate the number of samples assigned to each texture class combination. LUCAS-derived raster texture class were harmonized to fit the ECHO 7 texture classes.

	Clay loam	Clay	Sand	Sandy clay	Sandy loam	Silty	Silty clay
Clay loam	123	1	4	3	674	63	1
Clay	123	0	23	1	585	29	1
Sand	106		62		570	21	
Sandy clay	171		28	3	764	43	
Sandy loam	172	1	43	4	1348	80	
Silty	133		12	1	383	45	
Silty clay	198	1	2	1	503	81	0

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Supplementary material

Advancing soil monitoring through citizen science-generated data

Céline Laurent, Claudia Cappello, Timo Breure, Oleg Osychenko, Arwyn Jones,
Cristiano Ballabio, Panos Panagos, Tanja Mimmo

¹Competence Centre for Plant Health, Free University of Bolzano, Bolzano 39100, Italy
²European Commission, Joint Research Centre (JRC), Ispra, Italy
³Quanta Systems S.L., Badalona 08912, Spain
⁴Faculty of Agricultural, Environmental and Food Sciences, Free University of Bolzano, Bolzano 39100, Italy

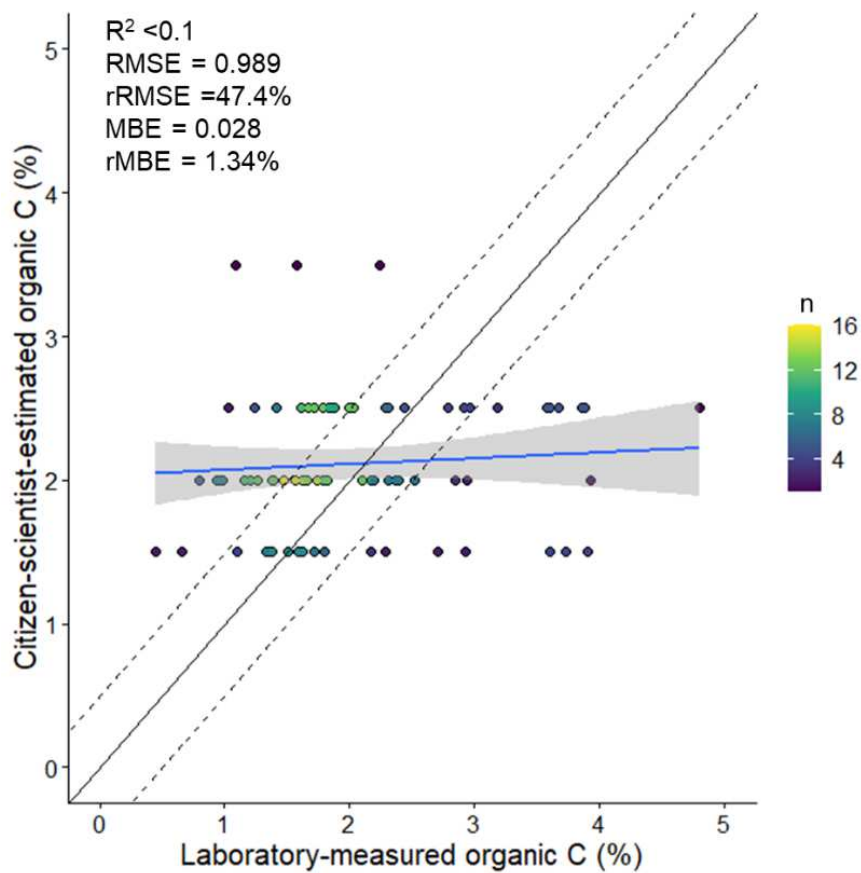


Figure S1: Soil organic carbon content (%) measured using citizen science method as a function of the same parameter measured using laboratory-based method. For each data point, the colour represents the number of samples (n), with the darker colour the fewer samples for each single value. The blue line, the thin solid line and the dashed lines represent the regression line with 95% confidence interval (grey shaded area), the 1:1 line and the ± 0.5 organic C %, respectively. The adjusted coefficient of the regression ($\text{adj } R^2$), the root mean square error (RMSE), the relative root mean square error (rRMSE), the mean bias error (MBE), and the relative mean bias error (rMBE) are calculated

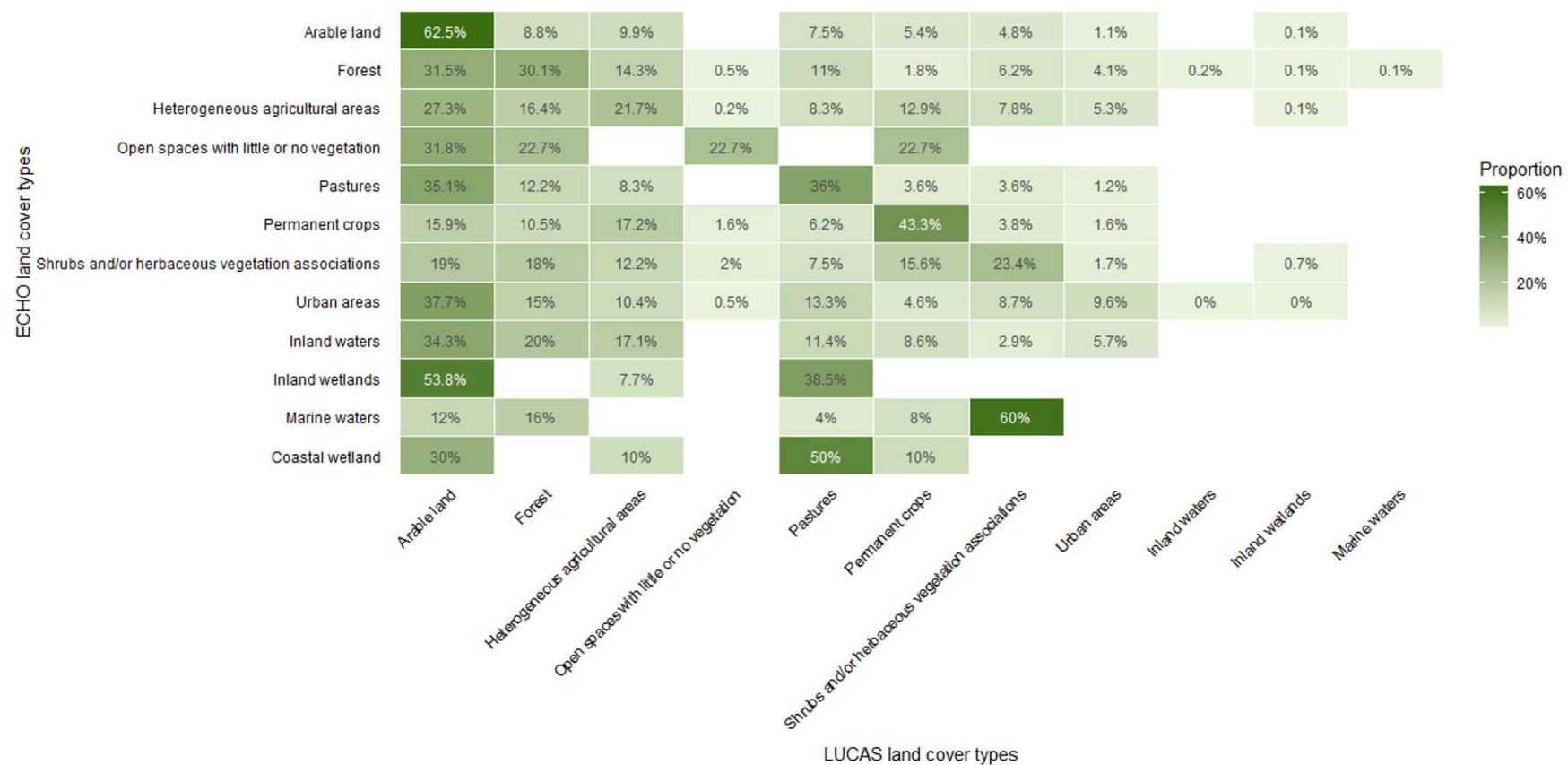


Figure S2: Proportion of LUCAS land cover types within each ECHO land cover types. Land cover types were derived from the CORINE Land Cover 2018 raster (100 m resolution). White cells indicate land cover combinations absent from the dataset.

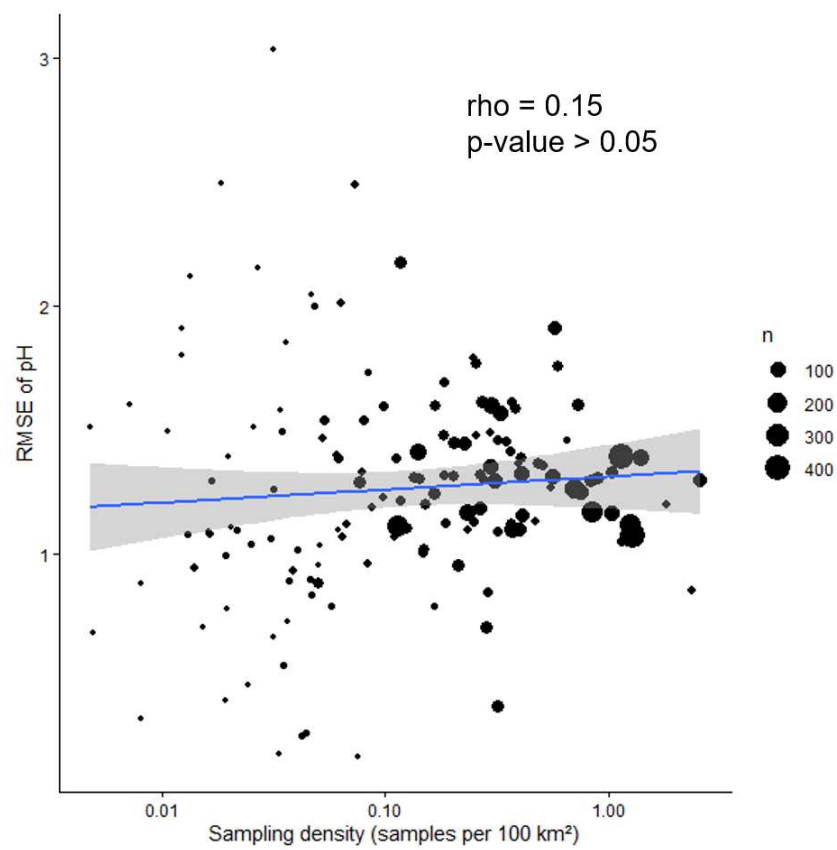


Figure S3: Root mean square error (RMSE) on pH measurements as a function of sampling density represented at NUT2 level. Each point represents a NUTS2 level.

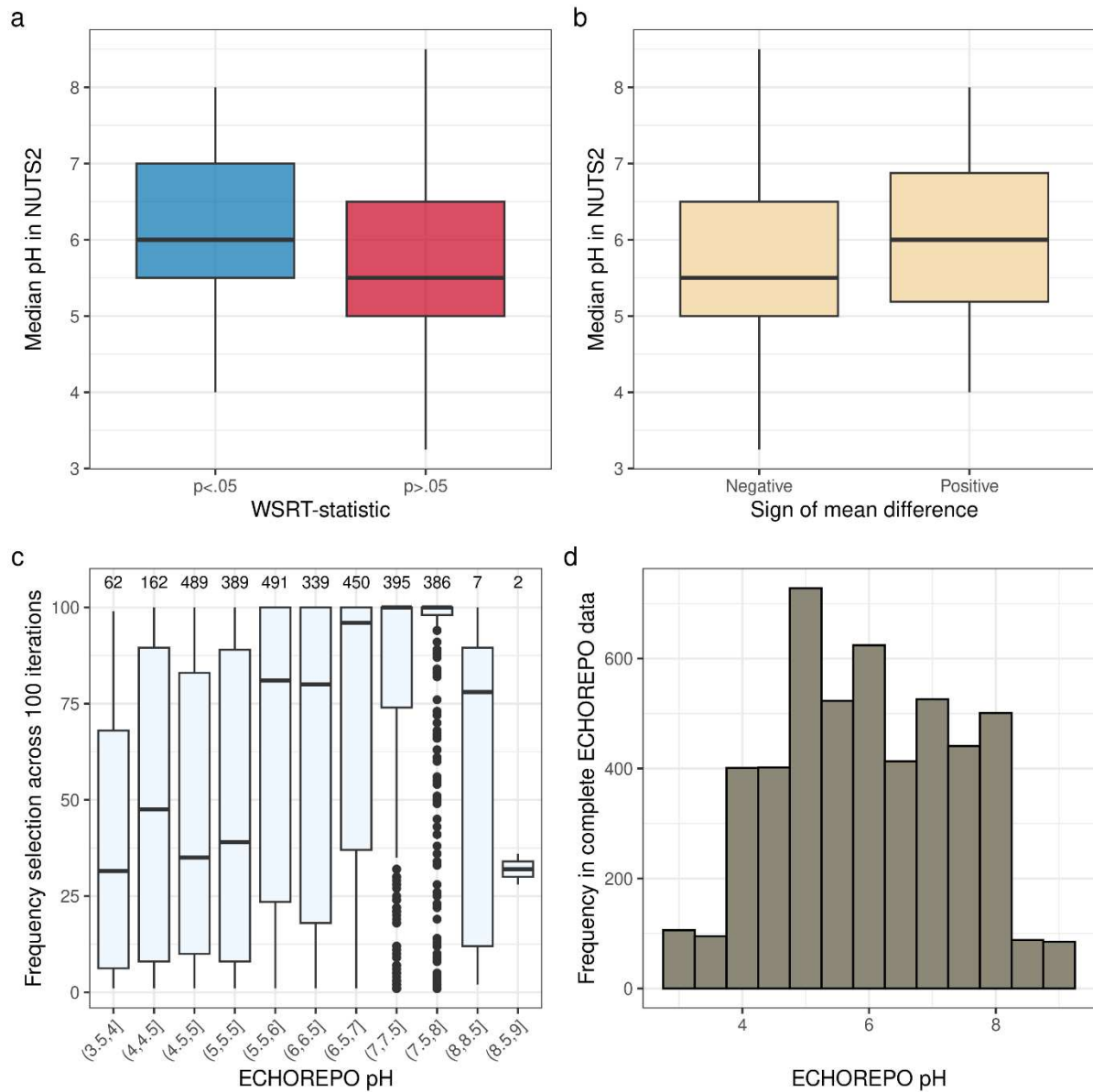


Figure S4: **a.** Median pH in NUTS2 region for a significant- or non-significant Wilcoxon signed-rank test (WSRT) statistic, **b.** Median pH in NUTS2 region for a negative- or positive mean difference between the CS and LUCAS implementation, **c.** The frequency in which a sample was selected across 100 iterations, where ECHOREPO pH values have been binned by a 0.5 interval. **d.** histogram of pH for the entire ECHOREPO dataset ($n = 6477$). All figures refer to where the CS implementation is for a sample size of ($N = 2000$) based on the IS-PIQ50-PIC weighting metric, except panel **d**.

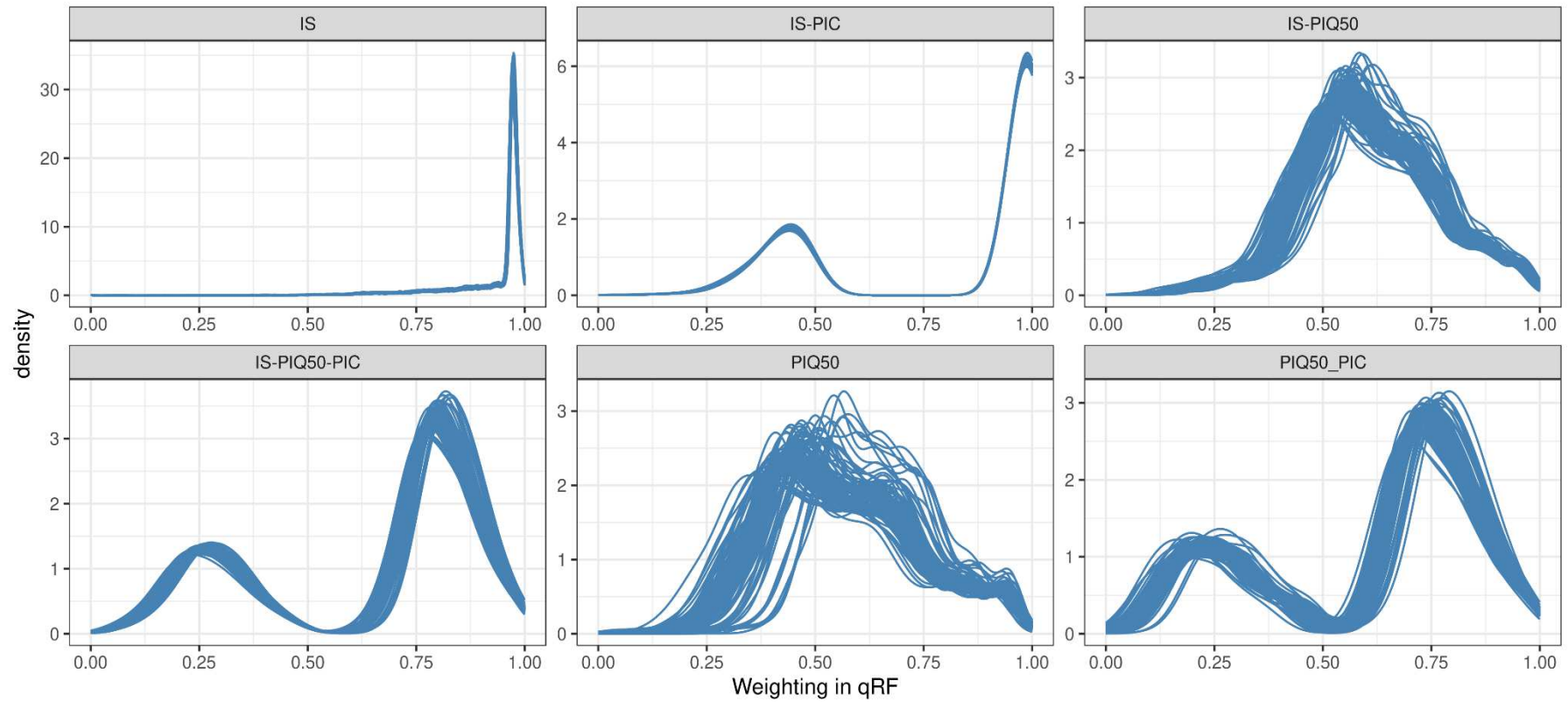


Figure S5: Kernel density distribution of the 90th prediction interval (95th quantile – 5th quantile) metrics that were used for weighting the ECHOREPO data to be included in the training dataset. IS is the interval score (Gneiting and Raftery, 2007). Pcov is a binary variable, whether the CS measurement fell within the PI or not. PI-Q50 is the prediction interval divided by the median ($(q_{95} - q_5)/q_{50}$).

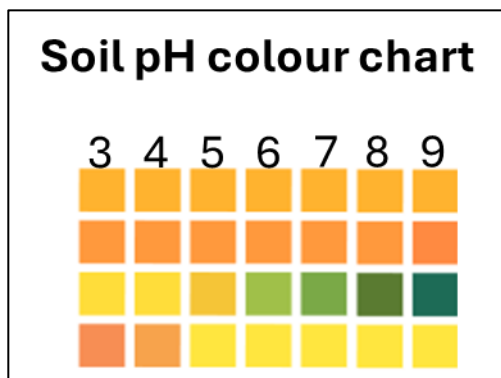


Figure S6: ECHO colour chart used in the ECHOsoil mobile application by citizen scientists to estimate soil pH, with pH values increasing from left to right, ranging from pH 3 (strongly acidic soil) up to pH 9 (strongly alkaline soil). Citizen scientists can then choose pH values, inside this pH range, separated by 0.5 pH unit path if the colour is not completely readable on the pH paper strip.

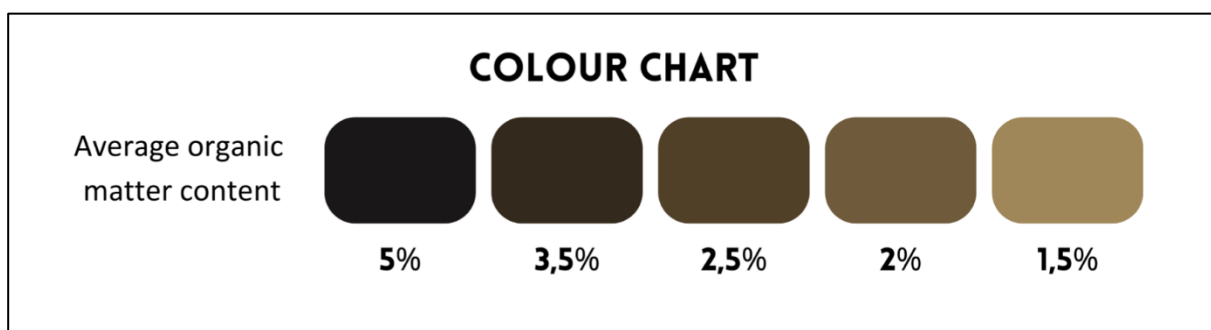


Figure S7: ECHO colour chart used by citizen scientists to estimate soil organic matter content from soil colour, with darkest shades indicating higher organic matter content (from 5% to 1.5%, from left to right).

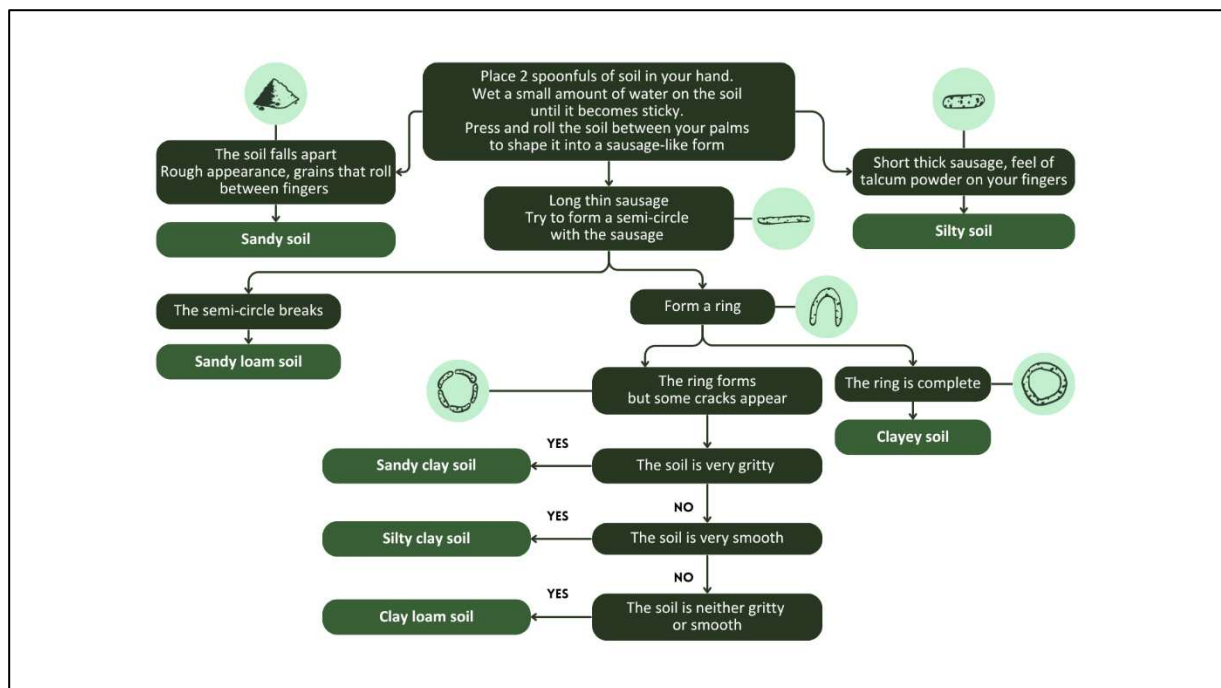


Figure S8: ECHO decision-making flowchart used by citizen scientists to determine soil texture, assessed by the “texture-by-feel” method (adapted from USDA soil quality guide).

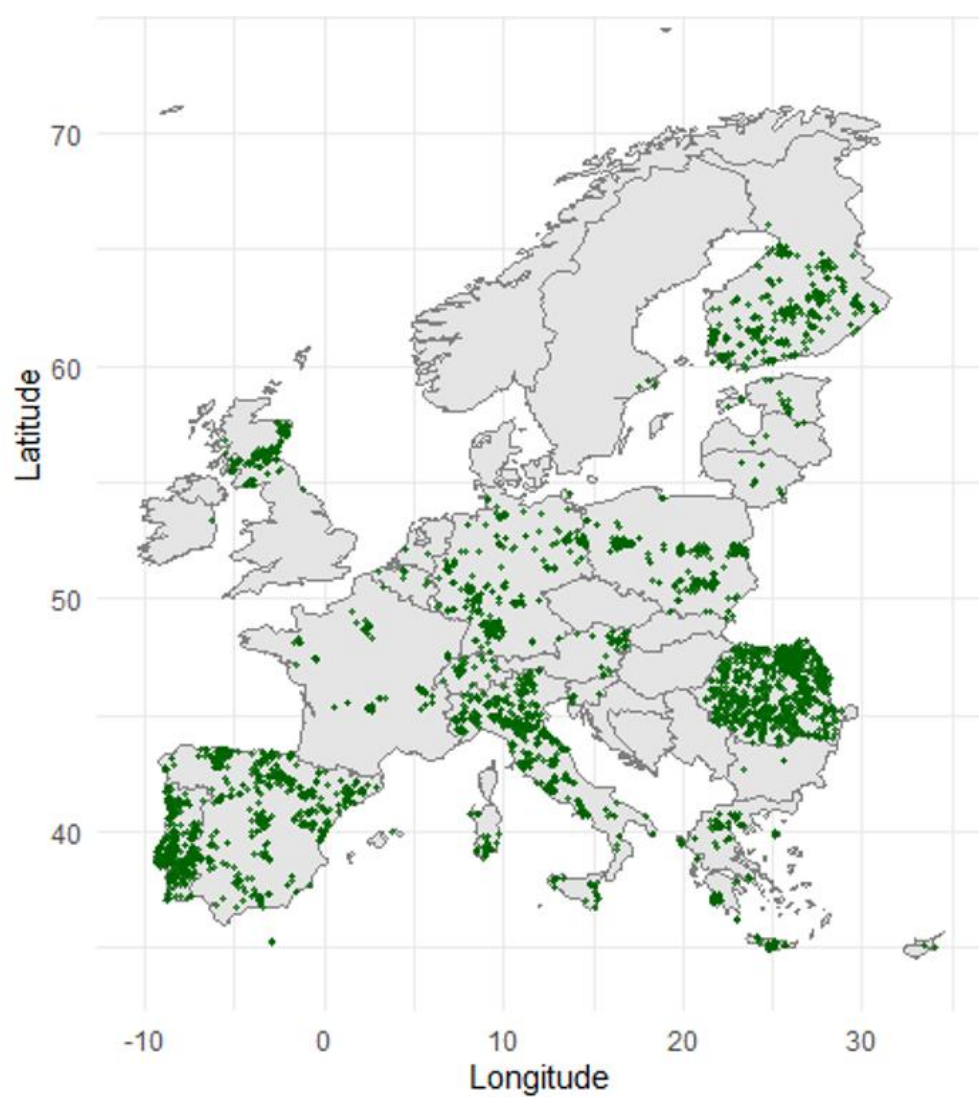


Figure S9: Map of ECHOREPO sampling locations

Table S1: Sampled countries and related number of ECHOREPO samples

Coutry	Number of ECHOREPO samples
Austria	110
Belgium	7
Bulgaria	9
Cyprus	2
Estonia	30
Finland	589
France	139
Germany	550
Greece	501
Ireland	1
Italy	968
Latvia	9
Lithuania	14
Luxembourg	2
Netherlands	3
Poland	668
Portugal	820
Romania	1059
Slovakia	2
Slovenia	28
Spain	878
Sweden	3
Switzerland	29
United Kingdom	308

Table S2: Major land cover types and related number of ECHOREPO samples

Land cover type	Number of ECHOREPO samples
Agricultural areas	2985
Artificial areas	2095
Forest and semi natural areas	1566
Wetlands	60
Water bodies	23

Table S3: Correspondence between LUCAS original soil texture classes and the harmonized classes used for comparison with the ECHOREPO dataset.

LUCAS original soil texture class	Harmonized class
Sand, Loamy sand	Sand
Sandy loam, Loam	Sandy loam
Sandy clay, Sandy clay-loam	Sandy clay
Silt, Silt loam, Silt clay-loam	Silt
Clay loam	Clay loam
Silty clay	Silty clay
Clay	Clay

Digital soil mapping methodology to integrate ECHOREPO with LUCAS

Covariates used in the integration of ECHOREPO with LUCAS data

Table S4 Covariates used for the digital soil mapping workflow to integrate ECHOREPO with LUCAS. This table corresponds to Table 1 in Ballabio et al. 2019, with the addition of the aridity index. All covariates were considered 500 meter resolution, see Ballabio et al. 2019 for further details.

Covariate code	Description
mir_PC <i>Abi</i>	Component <i>i</i> of PCA transformed of MODIS multitemporal Mean Infrared band for year 2009
nir_PC <i>Abi</i>	Component <i>i</i> of PCA transformed of MODIS multitemporal Near Infrared band for year 2009
red_PC <i>Abi</i>	Component <i>i</i> of PCA transformed of MODIS multitemporal Red band for year 2009
blue_PC <i>Abi</i>	Component <i>i</i> of PCA transformed of MODIS multitemporal blue band for year 2009
phen_MODIS <i>i</i>	Periodic component <i>i</i> of MODIS NDVI time series Fourier harmonic analysis
trend_MODIS <i>i</i>	Trend component <i>i</i> of MODIS NDVI time series Fourier harmonic analysis
tmax <i>i</i> _500	Average max temperature of month <i>i</i> from WorldClim
tmin <i>i</i> _500	Average min temperature of month <i>i</i> from WorldClim
prec <i>i</i> _500	Average precipitation of month <i>i</i> from WorldClim
bio <i>i</i> _500	Bioclimatic index <i>i</i> from WorldClim
ycoord	Latitude
Xcoord	Longitude
elevation	Elevation
valley height	Valley height index
gen_surface	Smoothed Elevation
Ls	RUSLE topographic factor (Slope Length and Steepness factor)
Aacn	Altitude above channel network
airflow_height	Effective Air Flow Heights (Böhner and AntoniĆ, 2009)
downsl_dist_grad	Downslope Distance Gradient (Hjerdt et al., 2004)
corine	Classes 1: <i>i</i> of CORINE land cover
parent.mat	Classes 1: <i>i</i> of ESDB parent material
aridity	Aridity index based on TerraClim

ECHOREPO sample selection for integration with LUCAS data

Some NUTS2 regions with CS data did not have associated LUCAS data, such as the Balearic islands, Switzerland and some small NUTS2, but we decided not to exclude these. However, due to lack of complete covariates we excluded: i.) 251 CS measurements, giving a total CS dataset of 6477 observations, ii.) 121 LUCAS measurements, giving a total LUCAS dataset of 12360 observations. For both CS and LUCAS, the locations were spread across various Member States.

Weighting schemes of citizen-science (CS) measurements

In our study, we tested the effect of selecting CS measurements by weighting schemes based on the 90th prediction interval metrics of the CS pH data from the LUCAS model (i.e. without adding the CS data). The weighting was implemented as follows:

Firstly, we calculated the following metrics: prediction-interval scaled by the maximum q50 value (PI-Q50):

Eq. 1

$$PIQ50_i = \frac{u_i - l_i}{\hat{y}_i}$$

Where u_i and l_i are the upper- and lower limits of the 90th prediction interval (q95 and q5, respectively) $\hat{y}_i$ the estimated median (q50). We also computed the interval score (IS) (Gneiting and Raftery, 2007):

Eq. 2

$$IS_i(\tau) = (u_i - l_i) + \frac{2}{1 - \tau} (y_i - u_i) \cdot \delta(y_i > u_i) + \frac{2}{1 - \tau} (l_i - y_i) \cdot \delta(y_i < l_i)$$

Where τ is the confidence level (90% PI) and u_i and l_i its upper- and lower limits, y_i is the CS measurement and δ is a boolean indicator function taking 1 if true and 0 if false. Lastly, we calculated whether the CS measurement lay within the prediction interval:

Eq. 3

$$PIC(\tau)_i = \delta(l_i < y_i < u_i)$$

Each of these were then standardized by applying the following equation:

Eq. 4

$$w_i = 1 - \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

such that lower values obtained higher weighting. The sum of each combination, to obtain a full-factorial comparison, was then also standardized by Eq. 5, without subtraction from 1 since they had been already inverted in the first step.

The prediction interval coverage probability (PICP), used not for the weighting but for the test-set evaluation, has a slightly different formulation:

Eq. 5

$$PICP(\tau) = \frac{1}{n} \sum_i^n \delta(l_i < y_i < u_i) / N$$

Where y_i is the LUCAS measurement and N the total sample size.

Implementation of the Wilcoxon signed rank test

After we calculated the total error on NUTS2 region level for each CS implementation, we compared the best performing CS implementation (IS-PIQ50-PIC). Specifically, we subtracted the error of the IS-PIQ50-PIC results with the error from the LUCAS implementation. Thus, negative differences meant that the CS implementation performed better and vice versa.

We then assessed the heteroskedasticity of differences and equal variance assumptions required for a regular t-test. Pairwise differences were evaluated using the Shapiro-Wilk test (Shapiro & Wilk, 1965) for each NUTS2 region. Homogeneity of variance between CS and LUCAS total errors was assessed using Levene's test (Levene, 1960), which is robust to departures from normality. Given that not all NUTS2 regions fulfilled these assumptions (Shapiro-Wilk $p < 0.05$ or Levene's test $p < 0.05$), we applied the Wilcoxon signed rank test (WSRT) to all NUTS2 regions for consistency. WSRT is a non-parametric alternative to the paired t-test (Wilcoxon, 1945).

Since we used random sampling to create the training/testing dataset, for each iteration: i.) different NUTS2 regions occurred in the testing dataset, ii.) a varying number of LUCAS test sites occurred in each NUTS2 region. Furthermore, some NUTS2 regions had a low number of LUCAS sites in general, which meant that in some cases a low number of sites would be used for calculating the prediction metrics. Therefore, we disregarded NUTS2 regions that had fewer than 2 LUCAS sites (10 out of 148). Additionally, we also regarded a NUTS2 region if it did not appear more than twice across the 100 iterations, since WSRT requires a sample size higher than two, but this did not occur.

To compare paired error distributions of CS and LUCAS implementations across NUTS2 regions, we applied a paired WSRT using the `wilcox.test` function in base R. To account for multiple comparisons across NUTS2 regions, p-values were adjusted using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995). Differences were considered statistically significant at $\alpha = 0.05$ (adjusted p-value). Mean differences were calculated for each NUTS2 region, and we reported the results with test statistics, 95% confidence intervals, and the adjusted p-values.

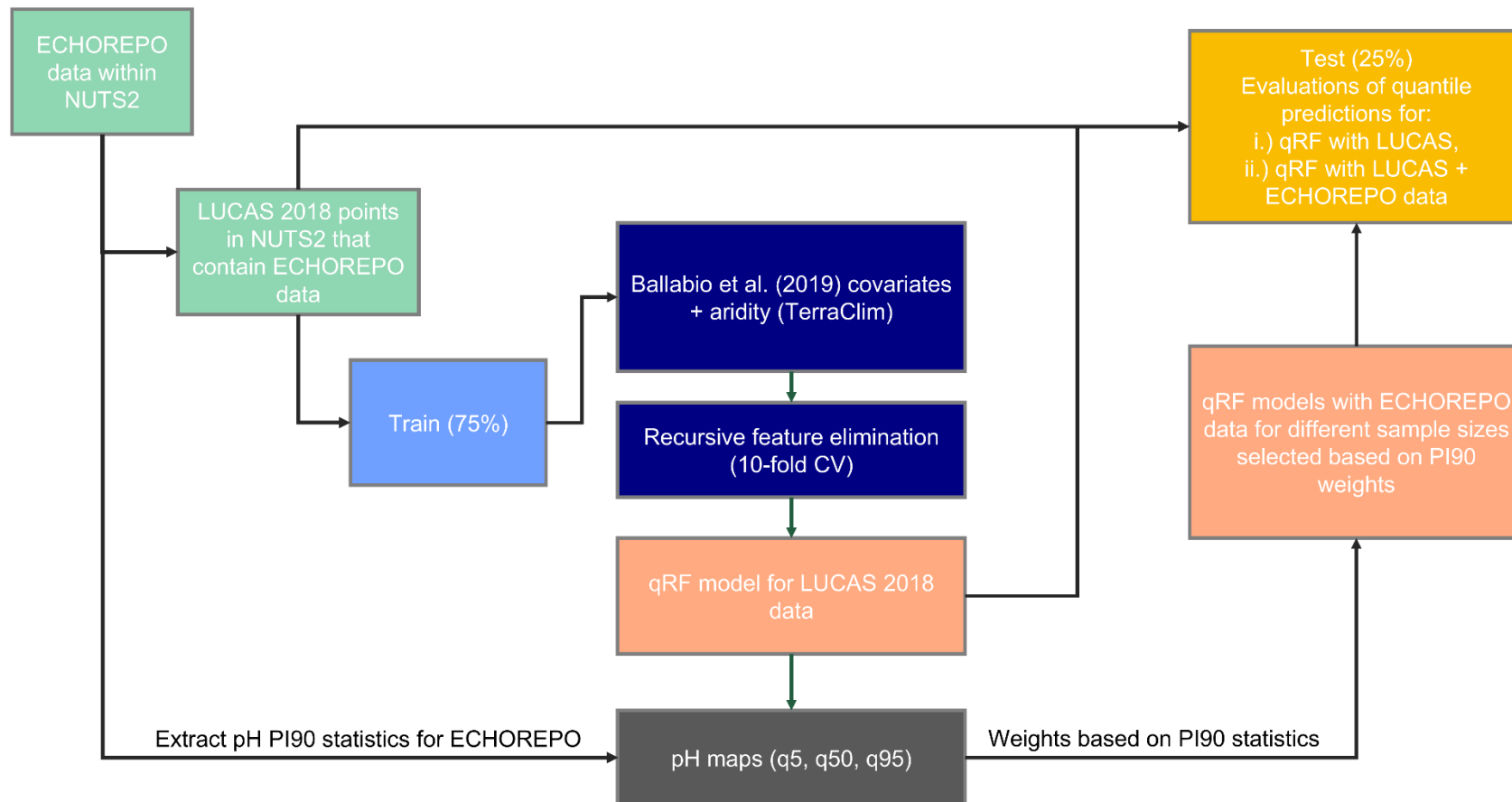


Figure S10: Flow-chart of the digital soil mapping procedure and tests to integrate the ECHOREPO citizen science data with the LUCAS dataset. qRF stands for quantile regression forest, PI90 stands for the 90th prediction interval, derived by subtracting the 95th quantile with the 5th quantile.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278–292). Stanford University Press.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591-611.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6), 80-83.