# Technical Note: A bootstrapped Loess regression approach for comparing soil depth profiles

**A. M. Keith*, P. Henrys, R. L. Rowe and N. P. McNamara**

{Centre for Ecology & Hydrology, Lancaster Environment Centre, Library Avenue, Bailrigg, Lancaster, LA1 4AP, United Kingdom}

*Correspondence to: A. M. Keith (ake@ceh.ac.uk)

## Abstract

Understanding the consequences of different land uses for the soil system is important to better inform decisions based on sustainability. The ability to assess change in soil properties, throughout the soil profile, is a critical step in this process. We present an approach to examine differences in soil depth profiles between land uses using bootstrapped Loess regressions (BLR). This non-parametric approach is data-driven, unconstrained by distributional model parameters and provides the ability to determine significant effects of land use at specific locations down a soil profile. We demonstrate an example of the BLR approach using data from a study examining the impacts of bioenergy land use change on soil organic carbon (SOC). While this straightforward non-parametric approach may be most useful in comparing SOC profiles between land uses, it can be applied to any soil property which has been measured at satisfactory resolution down the soil profile. It is hoped that further studies of land use and land management, based on new or existing data, can make use of this approach to examine differences in soil profiles.


## 1 Introduction

Understanding the consequences of different land uses for the soil system is important to better inform decisions based on sustainability (Foley et al., 2005; Haygarth and Ritz, 2009). The ability to assess change in soil properties effected by altered land use or management is therefore a critical step in this process. Greatest change is likely in the surface layers with factors such as tillage and plant inputs impacting the physical, chemical and biological properties of the soil. Many soil properties, however, will also be modified below this depth, particularly as time since land use change (LUC) increases (Popelau et al., 2011). It is therefore important that changes can be assessed below the topsoil and throughout the soil profile.

As a prime example, a number of studies, including global meta-analyses, have summarised the impacts of LUC on soil organic carbon (SOC) concentration and stocks (e.g. Guo and Gifford, 2002; Maquere et al., 2008; Laganière et al., 2010; Poeplau et al., 2011). SOC (*sensu* organic matter) is generally concentrated in the top 30 cm of the soil and so LUC is generally expected to have the greatest impact on SOC in these upper layers (Lorenz and Lal, 2005; Laganière et al., 2010). Even within this surface soil, however, the magnitude and sometimes direction of the effects of LUC on SOC can depend on the depth that is being considered (Guo and Gifford, 2002; Popelau et al., 2011). It is also becoming more evident that, in addition to there being a large proportion of total SOC stocks resident in the subsoil, important C dynamics may also occur deeper in the soil (Jobbágy and Jackson, 2000; Lorenz and Lal, 2005).

The turnover time of SOC generally increases with depth and hence the stabilisation of C may take place in deeper soil. Stabilisation pathways are likely through biochemical stabilisation driving reduced decomposition, by the inherent recalcitrance of root litter (e.g. lignins) and by physicochemical stabilisation (e.g. complexing with minerals and clay in subsoils)(Lorenz and Lal, 2005). Conversely, priming of the decomposition of older SOC may occur following LUC, especially with woody species (see Fontaine et al., 2007). This is particularly relevant for LUC to perennial vegetation or forest where deeper rooting plants are involved. For example, the root systems of perennial or tree species are likely to be more permanent and extensive in the subsoil, with a greater contribution of recalcitrant litter and potential priming down the soil profile (Fontaine et al., 2007). Altered land use or management may also impact the translocation of particulate and dissolved organic C likely to occur down the soil profile via effects on leaching. Such mechanisms may produce more complex relationships between soil depth and soil characteristics, and even discontinuous horizonation, rather than linear gradients.

## 2. Existing approaches to model soil depth profiles

Differences in SOC across transitions and soil depth profiles can be tested with both land use and depth included as fixed factors in an interaction model, and appropriate random terms to account for non-independence of depth increments within the same core and/or plots. There are, however, various potential modelling approaches that have been used to examine soil depth profiles including, for example, modified exponential decay (Maquere et al., 2008), depth distribution functions which utilise multiple regression (Indorante et al., 2013) and spline functions (Bishop et al., 1999; Malone et al., 2009; Wendt and Hauser, 2013). Another common method for non-linear modelling is the use of Generalised Additive Models (Hastie and Tibshirani, 1990).

Recent work modelling depth profiles has focussed on deriving parametric non-linear relationships between soil depth and the response of interest. Maquere et al. (2008) adopt a parametric form with modified exponential decay, whereas Myers et al. (2011) use an approach based on asymmetric peak functions. Whilst capturing the non-linear form of the soil depth profile, neither exponential decay nor polynomial methods adequately handles the associated uncertainty and hence confidence intervals, with the method in Maquere et al. (2011) assuming a *t*-distribution and the method in Myers et al. (2008) failing to produce confidence envelopes at all. Regression-based approaches similar to the popular GAM method have also been adopted using multiple covariates to account for any non-linearity (Indorante et al., 2013) and fitting cubic splines directly (Wendt and Hauser, 2013). However, the multiple regression approach assumed a normal distribution of the response variables, which is often not realised, and the cubic spline method presented by Wendt and Hauser does not provide any measure of uncertainty.

Non-linear relationships between SOC and soil depth across LUC transitions can also be incorporated by the inclusion of flexible splines (Bishop et al. 1999; Wood, 2003; Malone et al. 2009). In particular, the use of equal-area smoothing splines has long been considered as a beneficial approach to alleviate issues of modelling continuous soil depth functions using increment or horizon data (Bishop et al. 1999) and recent work has utilised the approach in the large-scale mapping of soil properties (Malone et al. 2009; Odgers et al. 2012; Adhikari et al. 2014). Equal-area spline functions consist of locally fitted quadratic functions tied together with knots at horizon boundaries (Malone et al. 2009), and the areas under/over the fitted curve optimised for equality in each horizon (Bishop et al. 1999). Confidence intervals and significance tests are, however, based upon the assumption that the response variable is drawn from the exponential family of distributions and inference is very sensitive to this assumption. Malone et al. (2009), in their study mapping continuous depth functions of SOC and water storage, highlighted the need for better estimation of uncertainty in such model outputs, suggesting the use of simulation and re-sampling approaches.

Simulation and re-sampling techniques avoid the necessity to assume a distributional form for the response variable in order to obtain confidence intervals and test hypotheses. Such approaches are rarely used to investigate soil depth relationships despite the often flawed assumptions made by the more commonly applied methods. Clifford et al. (2014) adopted a simulation routine from a master database to impute missing values and this clearly demonstrated another strength of the simulation approach, though they did not apply the method directly to test specific hypotheses relating to changes along the soil profile.

We sought to develop an approach which 1) would be able to compare and test for significant differences between potentially non-linear depth profiles of land uses (or across land use transitions), 2) did not need to meet any parametric distribution assumptions given that individual datapoints in soil datasets are typically non-independent (i.e. vertically or horizontally nested measurements) and 3) would be generally applicable regardless of specific contexts of land use and soil type. Below, we describe the resulting non-parametric approach and provide an example comparing SOC depth profiles across a land use transition.

## 3. A bootstrapped Loess regression (BLR) approach

The developed approach combines bootstrapped resampling of data with local least squares-based polynomial smoothing (Loess) regression. Consequently, this non-parametric method benefits from being data-driven and unconstrained by distributional form or rigid model parameterisation. Like spline approaches (Malone et al. 2009; Wendt and Hauser, 2013), it doesn't assume constant values for soil layers or horizons. Such a non-parametric approach is highly suitable where data are non-independent. This is particularly applicable in soil profiles where measurements made in depth increments down a soil profile may be correlated and even more relevant where data are cumulative (e.g. cumulative C stocks). It is also appropriate where soil cores have been sampled using a nested spatial design with multiple cores taken from within plots.

The BLR approach is intended to make use of soil data which has been measured at fixed-depth intervals down the soil profile at a generally high resolution, or at least at a resolution satisfactory for the purposes of an assessment. The vertical sampling resolution is not limited to any specific depth interval (e.g. 10 cm increments) but clearly a greater, and regular, resolution provides more detailed information on potential differences and their specific location in the soil profile. Low sample sizes will affect the amount of smoothing that can be done by the Loess algorithm. As the algorithm fits polynomial regressions within local neighbourhoods, the definition and size of the neighbourhood determines the smoothness and sensitivity of the fitted regression line. Typically a minimum of 3 observations per neighbourhood would be required.

The initial dataset comprises all data for the soil variable of interest from the two land uses ($LU_1 \cup LU_2 = LU_{ALL}$) which are to be compared, with the associated depth and/or soil mass as reference. A subset is then created containing only data from the 'second' land use ($LU_2$). In theory, it doesn't matter which of the land uses are subsetted for $LU_2$ but one may be more intuitive given the direction of a specific land use transition. It is also useful to plot the data to determine whether the datasets contain outliers that may need to be excluded before bootstrapping to prevent skewing the Loess regression. For cumulative mass-based data, if datapoints from the bottom depths of either $LU_1$ or $LU_2$ are at distinctly greater cumulative masses than others, these could also be trimmed so that the comparison is made to the same approximate lower bounds of the reference. Using a large number of bootstrap samples, however, should negate the need for extensive data cleansing prior to analysis.

The combined data ($LU_{ALL}$) is re-sampled by bootstrap with replacement, with the number of data-points resampled equal to the number of data-points in $LU_2$. This is repeated n=1000 times. Each bootstrapped set of data are then modelled using Loess regression and these regressions are used to generate 95% confidence intervals around a modelled soil depth profile by taking pointwise percentiles at each depth. As each sub-sample is taken from the union of the two land uses, this confidence interval (or confidence envelope) represents the null hypothesis that there is no difference between the $LU_1$ and $LU_2$. The data from only $LU_2$ is then modelled using Loess regression; if the modelled line for the $LU_2$ profile sits outside the confidence envelope of the null hypothesis it can be inferred that the soil variable is significantly different between $LU_1$ and $LU_2$ at that particular point in the profile. Overall P values for the difference between depth profiles can be obtained by taking normalised test statistics across the full set of bootstrap samples and taking the percentile of these values corresponding to the same statistic obtained from the $LU_2$ data. This is a similar approach to that adopted in the spatial statistics literature when analysing K functions under resampling as demonstrated in Diggle et al. (2008) and Henrys and Brown (2009) for example.

This relatively straightforward non-parametric method may be most useful in comparing SOC profiles between land uses, but it can be applied to any soil property which has been measured at satisfactory resolution down the soil profile. Many of these other properties measured in soil (e.g. bulk density, pH, root biomass) can vary in a non-linear fashion down the soil profile, with potential horizonation. The effects of land use change are typically examined using either a paired-site or chronosequence approach. These assume that each paired or chronosequence site only differs in their age or, for example, time since disturbance and have comparable biotic and abiotic histories (Laganière et al., 2010). While this BLR method benefits from being unconstrained by assumptions of parametric methods, it must still satisfy the assumptions of the paired-site and chronosequence approaches, particularly if space-for-time substitution is used (Indorante et al., 2013). Here, we provide an example comparing SOC depth profiles between land uses. The approach is, however, not limited to comparing soil depth profiles between land uses. It could also be usefully adopted to examine, for example, depth functions in lake systems or to compare temporal trajectories in soil metrics between experimental treatments.

## 4. Applying a BLR approach - an example of bioenergy land use change

The bootstrap re-sampling and Loess regression used to test differences between soil profiles was conducted using the R statistical programming language (R Core Team, 2015). Example code to demonstrate the BLR approach using real data are available via http://doi.org/10/f3jp5d (Keith et al., 2015). These data are from a study examining the impacts of bioenergy LUC on SOC in the UK (Rowe et al., 2016). A LUC transition from arable to Short-Rotation Coppice (SRC) willow was selected and the data were separated into subsets of those from each component of the transition (i.e. arable and SRC willow samples) before analysis. Data on SOC concentration (expressed as a percentage), cumulative SOC stock and cumulative dry soil mass were derived at 10 cm increments to 1 m depth in order to construct fixed-depth profiles of SOC concentration (Figure 1A; Figure 2A and C) and mass-based depth profiles of SOC stocks (i.e. the relationship between soil mass and SOC *sensu* Gifford and Roderick, 2003) (Figure 1B; Figure 2B and D). Cumulative soil mass was used because measured SOC stock in small fixed-depth increments (as was required in this study) may not be directly comparable across LUC transitions, due to potential variation in bulk densities and any compression or expansion introduced through sampling (e.g. Gál et al., 2007). An approach using soil mass as the independent variable overcomes this issue more generally because profiles can be directly compared at a particular reference soil mass (Gifford and Roderick, 2003; Wendt and Hauser, 2013). Gifford and Roderick (2003) suggest a reference dry soil mass of 4000 t $ha^{-1}$ and 12000 t $ha^{-1}$ may be used to approximate sampling to 30 cm and 1 m depth in agricultural systems, respectively. This is not an issue when examining SOC concentration, as these data are not directly influenced by core volume and apparent bulk density.

There was generally a good fit between observed and modelled data, with all modelled means well within a standard error of the observed means in each depth increment, and the majority very close to the actual observed mean (Figure 1). The poorest fits appeared to be for SOC concentration around the plough layer of the arable land use (20-40 cm; Figure 1A) and in the upper layers of the SRC willow land use (0-20 cm; Figure 1C). The RMSE values for the depth profiles were 0.037, 0.487, 0.028 and 0.929 for arable SOC concentration, arable SOC stock, SRC willow SOC concentration and SRC willow SOC stock, respectively.

Individual datapoints for each land use, the confidence envelope of the null hypothesis and the modelled profile for the SRC willow were plotted following BLR (Figure 2). Where the modelled line sits outside the confidence envelope it can be inferred whether there are significant effects of land use in the soil profile, at either a particular depth or references soil mass. In Figure 2A, the SOC concentration is significantly greater under SRC willow compared with arable at 10 cm and 20 cm, where the modelled line sits to the right of the confidence envelope. The modelled line sits within the confidence envelope between 40 cm – 100 cm and so there is no significant difference (Figure 2A). Nevertheless, the two depth profiles are significantly different overall ($P < 0.01$). The depth profile of SOC concentration is reflected in the cumulative SOC stock profile, with the modelled line for SRC willow moving further from the confidence envelope up to approximately 5000 t $ha^{-1}$ (Figure 2B). The difference in cumulative SOC stock between arable and SRC willow is maintained to 100 cm and, consequently, is significantly different down the whole soil profile ($P < 0.01$; Figure 2B).

## 5. Conclusions

We modelled soil profiles and tested differences in soil characteristics between land use or land management using a non-parametric approach combining bootstrap sampling and Loess regression. The development of this approach was driven by a need for a flexible method which could compare potential non-linear relationships between land uses (or across land use transitions) and would not be constrained to specific contexts. While there are several other methods which can be used to model non-linear relationships in soil depth profiles, the BLR approach is flexible because it is data-driven and does not need to meet any distributional assumptions. The confidence envelopes obtained are robust to miss-specification of the error distribution and provide clear inspection of significant differences across the full depth profile. There can be issues of model fit when profiles are discontinuous or change abruptly. This is not exclusive to the BLR approach though and it also affects equal-area spline models (see Odgers et al., 2012). It has been proposed that the use of pseudo-horizons may help towards overcoming this challenge (Malone et al. 2009; Odgers et al., 2012). We acknowledge that in some circumstances the equal area spline functions are a viable alternative to Loess regression for producing a fitted profile. This could, however, easily be incorporated into the non-parametric estimation and bootstrapping framework that we present here.

Sampling to depth and increasing the resolution of depth increments can provide useful profiles or 'fingerprints' of soil properties under different land uses and soil types. In particular, assessment of SOC to depth, and determining the response of SOC to land use change (LUC) or land management change is essential to understand the sustainability of different soil use options. This may be particularly important for land-use transitions to perennial crops, which have deeper and more permanent rooting systems that may influence the C balance deeper in the subsoil via priming of decomposition, C stabilisation or translocation. The BLR approach can be, however, applied to any soil property of interest giving the ability to assess land use effects at any point down the soil profile. Being data-driven and flexible, it is hoped that further studies of land use and land management, based on new or existing data, can make use of this approach to examine differences in soil profiles.

*Author contributions.* A. M. Keith and R. L. Rowe conducted sampling and created the data. A. M. Keith and P. Henrys developed the statistical approach. All authors contributed to preparation of the manuscript.

*Acknowledgement.* This work was supported by the ELUM (Ecosystem Land Use Modelling & Soil Carbon GHG Flux Trial) project, which was commissioned and funded by the Energy Technologies Institute (ETI).

## References

Adhikari, K., Hartemink, A. E., Minasny, B., Bou Kheir, R., Greve, M. B., and Greve, M. H.: Digital mapping of soil organic carbon contents and stocks in Denmark, PLoS One 9, e105519, 2014.

Bishop, T. F. A., McBratney, A. B., and Laslett, G. M.: Modelling soil attribute depth functions with equal-area quadratic smoothing splines, Geoderma, 91, 27-45, 1999.

Clifford, D., Dobbie, M. J., and Searle, R.: Non-parametric imputation of properties for soil profiles with sparse observations, Geoderma, 232-234, 10-18, 2014.

Diggle, P. J., Gómez-Rubio, V., Brown, P. E., Chetwynd, A. G., and Gooding, S.: Second-Order Analysis of Inhomogeneous Spatial Point Processes Using Case–Control Data, Biometrics, 63, 550-557, 2007.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowksi, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., Prentice, I. C., Ramankutty, N., and Snyder, P. K.: Global consequences of land use, Science, 309, 570-574, 2005.

Fontaine, S., Barot, S., Barre, P., Bdioui, N., Mary, B., and Rumpel, C.: Stability of organic carbon in deep soil layers controlled by fresh carbon supply, Nature, 450, 277-280, 2007.

Gál, A., Vyn, T. J., Michéli, E., Kladivko, E. J., and McFee, W. W.: Soil carbon and nitrogen accumulation with long-term no-till versus mouldboard plowing overestimated with tilled-zone sampling depths, Soil Till. Res., 96, 42-51, 2007.

Gifford, R. M., and Roderick, M. L.: Soil carbon stocks and bulk density: spatial or cumulative mass coordinates as a basis of expression? Glob. Change Biol., 9, 1507-1514, 2003.

Guo, L. B., and Gifford, R. M.: Soil carbon stocks and land use change: a meta analysis, Glob. Change Biol., 8, 345-360, 2002.

Hastie, T. J., and Tibshirani, R. J.: Generalized additive models (Vol. 43), CRC Press, 1990.

Haygarth, P. M., and Ritz, K.: The future of soils and land use in the UK: Soil systems for the provision of land-based ecosystem services, Land Use Policy, 26S, S187-S197, 2009.

Henrys, P. A., and Brown, P. E.: Inference for clustered inhomogeneous spatial point processes, Biometrics, 65, 423-430, 2009.

Indorante, S. J., Kabrick, J. M., Lee, B. D., and Maatta, J. M.: Quantifying Soil Profile Change Caused by Land Use in Central Missouri Loess Hillslopes, Soil Sci. Soc. Am. J., 78, 225-237, 2013.

Jobbágy, E. G., and Jackson, R. B.: The vertical distribution of soil organic carbon and its relation to climate and vegetation, Ecol. Appl., 10, 423-436, 2000.

Keith, A.M., Henrys, P. A., Rowe, R. L., and McNamara, N. P.: Bootstrapped local regression (LOESS) for soil depth profile comparison. NERC Environmental Information Data Centre, doi: 10.5285/d4f92cd8-43e8-49e4-8f9e-efcc0e3b2478, 2015.

Laganière, J., Angers, D. A., and Parè, D.: Carbon accumulation in agricultural soils after afforestation: a meta-analysis, Glob. Change Biol., 16, 439-453, 2010.

Lorenz, K., and Lal, R.: The depth distribution of soil organic carbon in relation to land use and management and the potential of carbon sequestration in subsoil horizons. Adv. Agron., 88, 35-66, 2005.

Malone, B. P., McBratney, A. B., Minasny, B., and Laslett, G. M.: Mapping continuous depth functions of soil carbon storage and available water capacity, Geoderma, 154, 138-152, 2009.

Maquere, V., Laclau, J. P., Bernoux, M., Saint-Andre, L., Gonçalves, J. L. M., Cerri, C. C., Piccolo, M. C., and Ranger, J.: Influence of land use (savanna, pasture, *Eucalyptus* plantations) on soil carbon and nitrogen stocks in Brazil, Eur. J. Soil Sci., 59, 863-877, 2008.

Myers, D. B., Kitchen, N. R., Sudduth, K. A., Miles, R. J., Sadler, E. J., and Grunwald, S.: Peak functions for modelling high resolution soil profile data, Geoderma, 166, 74-83, 2011.

Odgers, N. P., Libohova, Z., and Thompson, J. A.: Equal-area spline functions applied to a legacy soil database to create weighed-means maps of soil organic carbon at a continental scale, Geoderma 189-190, 153-163, 2012.

Poeplau, C., Don, A., Vesterdal, L., Leifeld, J., Van Wesemael, B., Schumacher, J., and Gensior, A.: Temporal dynamics of soil organic carbon after land-use change in the temperate zone – carbon response functions as a model approach, Glob. Change Biol., 17, 2415-2427, 2011.

R Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org, 2015.

Rowe, R. L., Keith, A. M., Elias, D., Dondini, M., Smith, P., Oxley, J., and McNamara, N. P.: Initial soil C and land use history determine, Glob. Change Biol. Bioenergy Early View Online, DOI: 10.1111/gcbb.12311, 2016.

Wendt, J. W., and Hauser, S.: An equivalent soil mass procedure for monitoring soil organic carbon in multiple soil layers, Eur. J. Soil Sci., 64, 58-65, 2013.

Wood, S. N.: Thin-plate regression splines. J. Roy. Stat. Soc. Ser. B, 65, 95-114, 2003.

## Figures

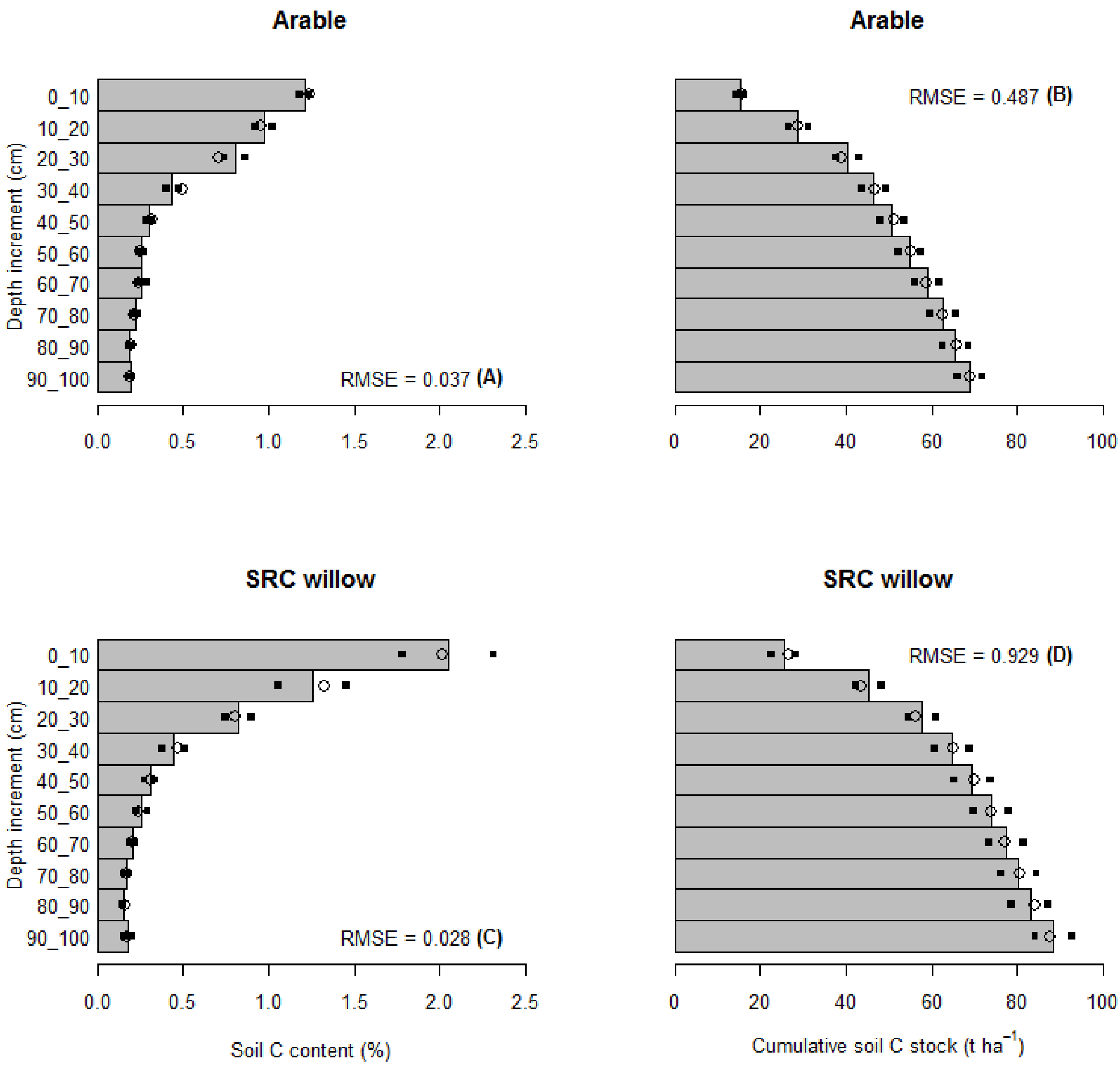


Figure 1. Soil carbon concentration (A,C) and cumulative carbon stock (B,D) of arable (A,B) and SRC willow (C,D) land uses in 10 cm depth increments to 1 m. Bars represent observed means, squares represent standard error of observed means, open circles represent modelled means. Root mean square error (RMSE) calculated for the depth profile using means of observed and modelled data from the ten depth increments.

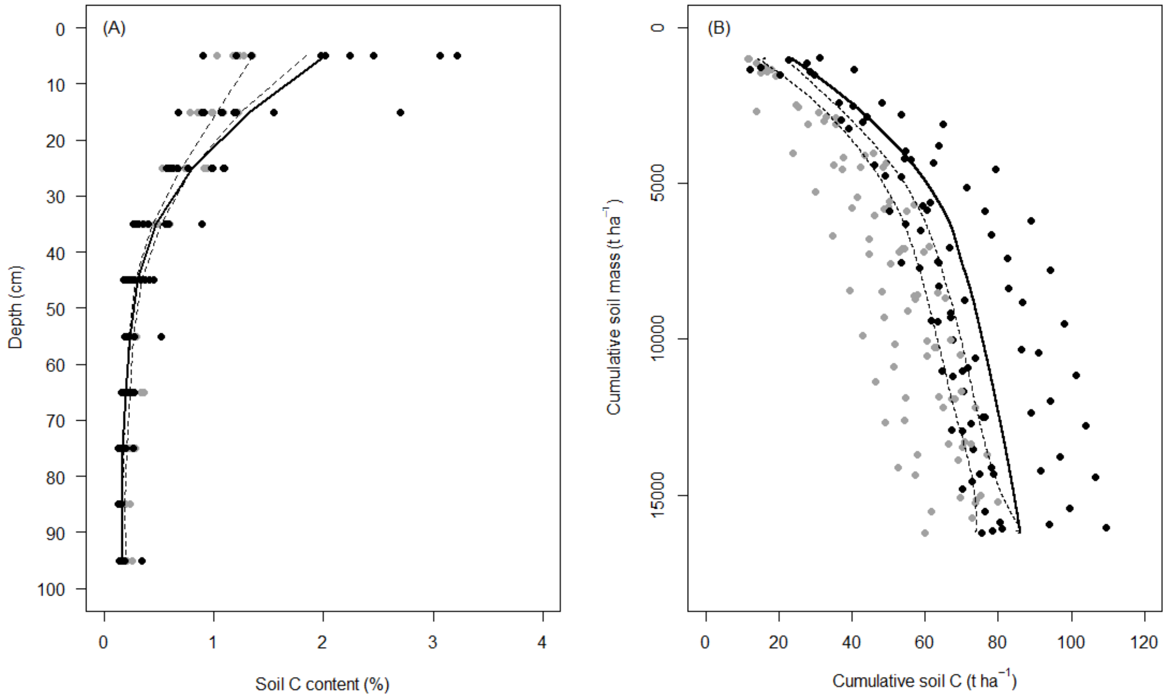


Figure 2. Difference in profiles of (A) Soil C concentration as a function of sampling depth and (B) Cumulative soil C stock as a function of soil mass. Depth represents values of samples from 10 cm increments. Grey and black symbols represent SRC willow and arable data-points, respectively. Dashed lines represent upper and lower bounds of 95% confidence intervals from bootstrapped (n = 1000) Loess regressions of combined arable and SRC willow data; solid lines represents Loess regression of percent C and cumulative soil C in SRC willow only, if this line sits outside the confidence interval it can be inferred that arable and SRC willow are significantly different.