European Commission

Joint Research Centre

The European Commission's in-house science service

www.jrc.ec.europa.eu

Serving society
Stimulating innovation
Supporting legislation

GlobalSoilMap.net at Jrc: data sources and ongoing activities

Cristiano Ballabio

European Commission

Institute for Environment and Sustainability

DG-JRC

Ispra

A consideration about Europe and DSM

- Europe (including Russia west of Urals) is slightly larger than Australia
- However, while Australia is a single country, at present Europe is composed of 50 countries
- Each country has a different soil survey program, different soil sampling designs and different analytical procedures
- To make things more complicate, in many countries soil mapping is managed at regional level (NUTS 1 or 2) by local agencies
- This fragmentation makes very difficult the mapping task as different datasets from different countries/regions might not be comparable and data might not be freely available

A consideration about Europe and DSM

- Although some data is available at EU level most of the data is still collected and managed at country or regional level

- Europe is probably one of the areas with the highest density of soil samples/profiles… unfortunately, is also one of the places where getting data might be very difficult…

Part one: the data being used

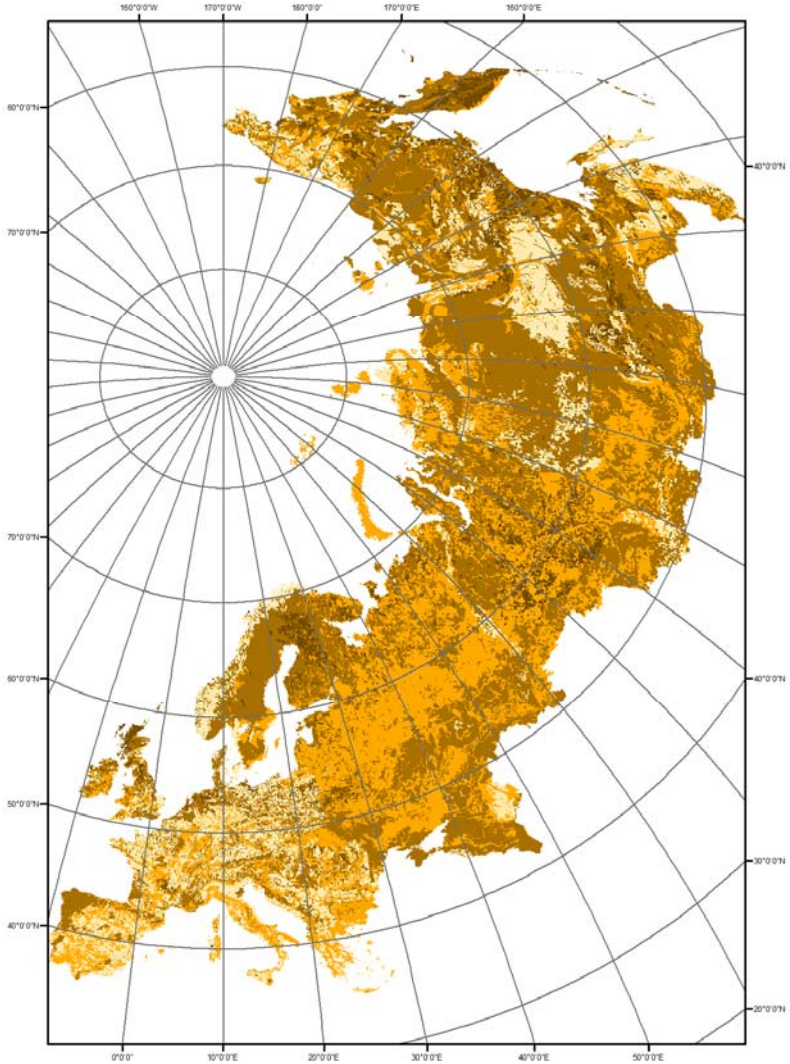
Available EU scale datasets

- Several EU wide datasets are available and can be further subdivided in:
- Georeferenced Point data
- Legacy Soil data (mostly maps)

Legacy data

- This is essentially Soil Geographical Database of Eurasia (SGDB) at a 1:1000000 scale

- It consists of a polygon map of soil types which follows the terminology of the F.A.O. legend for the Soil Map of the World at Scale 1:5000000

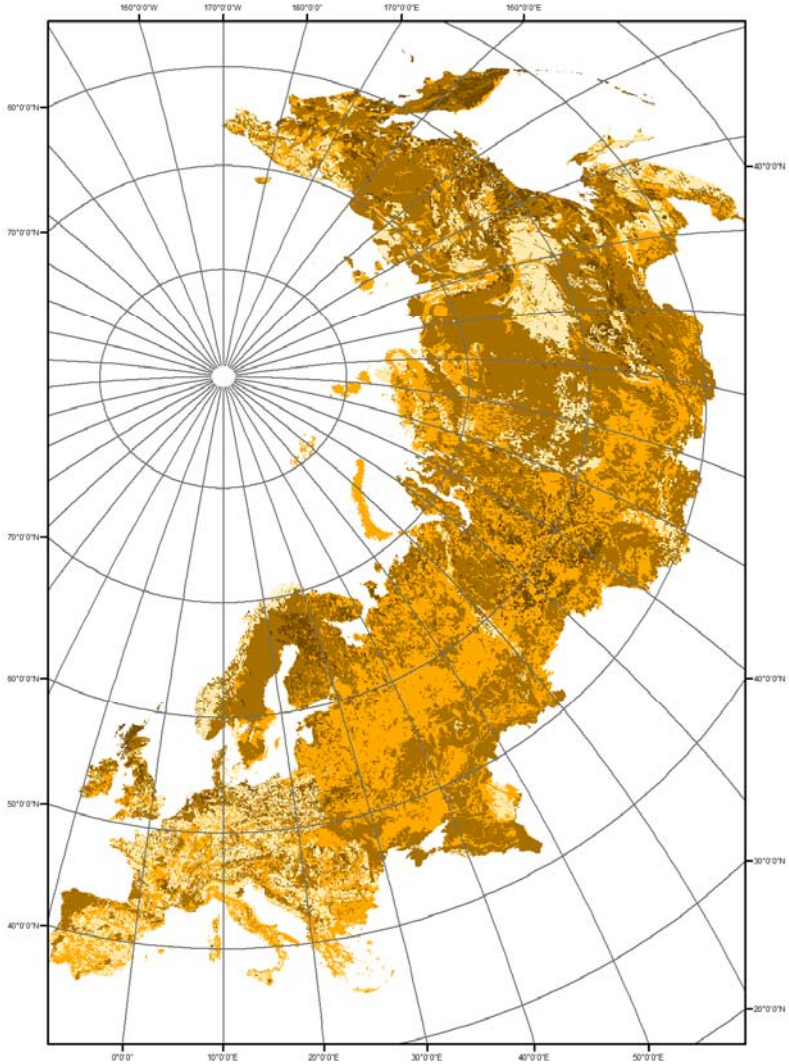


European Soil Database

- The European Soil Database (ESDB) at Scale 1:1,000,000 is part of the European Soil Information System (EUSIS)
- It is a simplified representation of the diversity and spatial variability of the soil coverage
- The FAO terminology has been refined and adapted to take account of the specificities of the landscapes in Eurasia. It is itself founded on the distinction of the main pedogenetic processes leading to soil differentiation: brunification, lessivage, podzolisation, hydromorphy

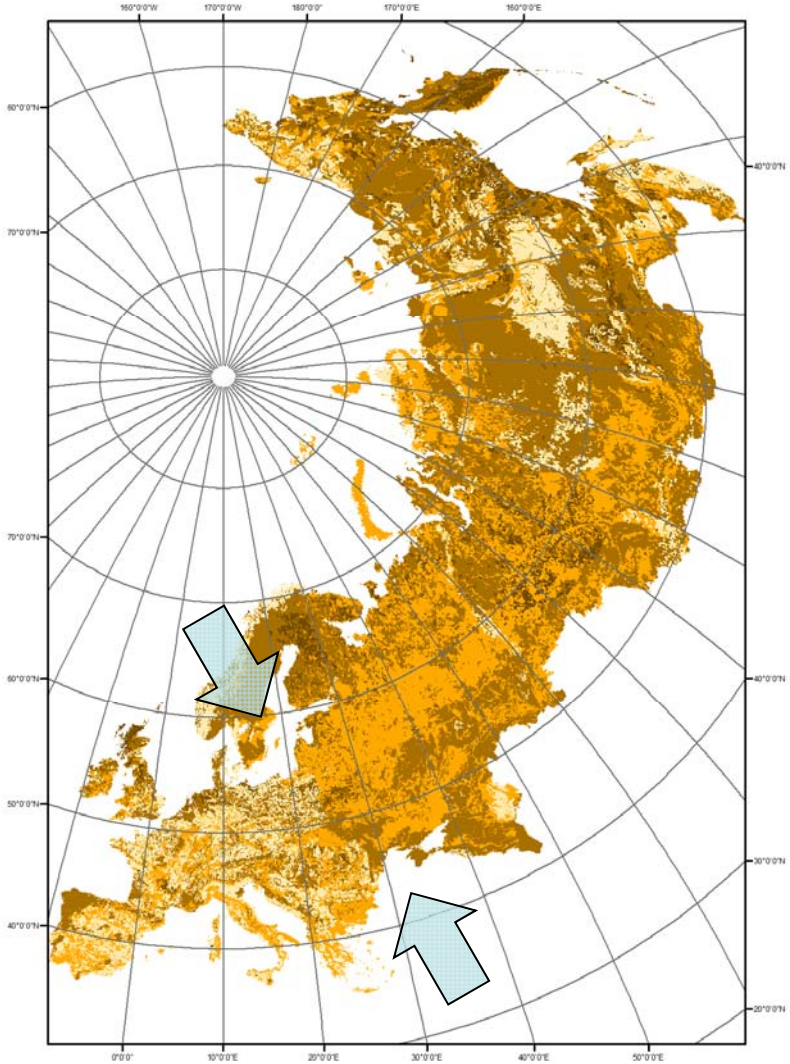
ESDB

- The actual data about soil features in the ESDB is rather scant
- Soil properties were estimated over large areas by expert judgement rather than measured on local soil samples
- A collection of classified soil properties for each soil unit is also present (i.e. ranked classes of Soil Organic Carbon)



ESDB harmonization

- In spite of being "harmonized" the SGDB shows some peculiar features…
- For instance the Soil Organic Carbon content changes in a rather abrupt way at the political border between Europe and Russia…

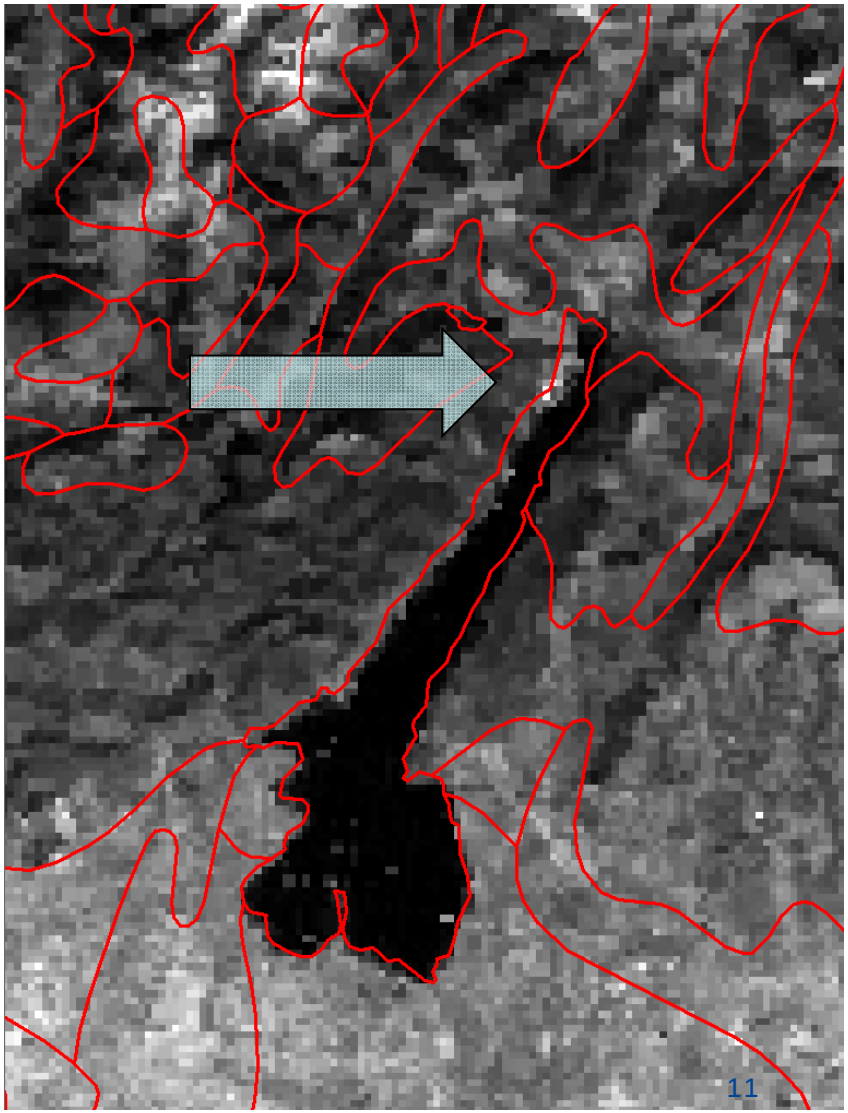


ESDB spatial accuracy

- Moreover, the usability of the Soil Map of Eurasia for DSM is somewhat questionable as many soil units do not follow terrain features properly

- Even at coarse resolution (i.e. 500m resolution MODIS) the shift is quite relevant

- Using fuzzy boundaries might help in solving this issue



Point data: LUCAS data

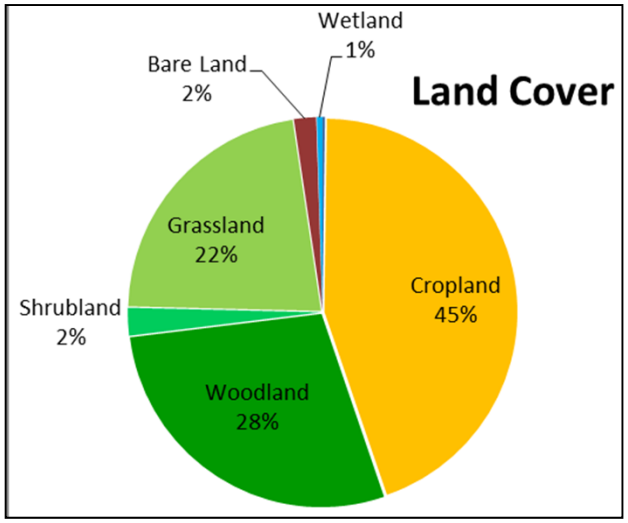
Selection of point, based on LUCAS (**L**and **U**se/**C**over **A**rea frame statistical **S**urvey) master grid (2 x 2 km), as a function of:

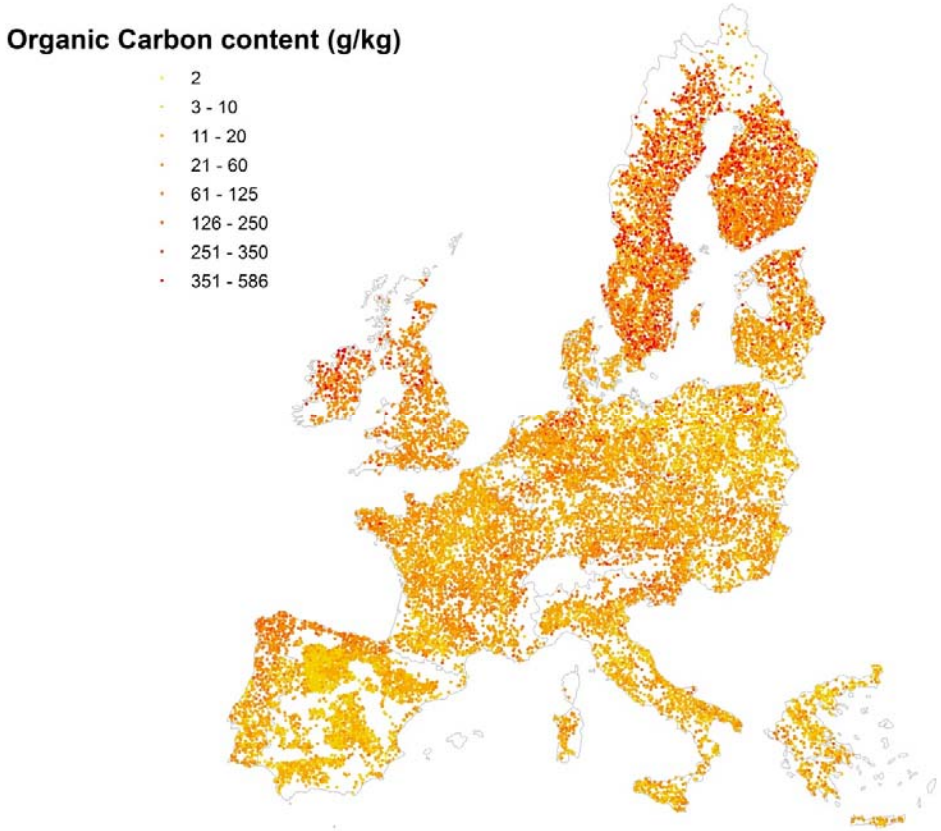
- country
- land use
- physiography
  - slope
  - aspect
  - elevation
  - slope
  - curvature, etc.

10 % = ~22,000 sampling points (triplets) selected for topsoil sampling

LUCAS data

LUCAS sampling excluded areas above 1000m of elevation and was mostly aimed at estimating topsoil properties in agricultural areas



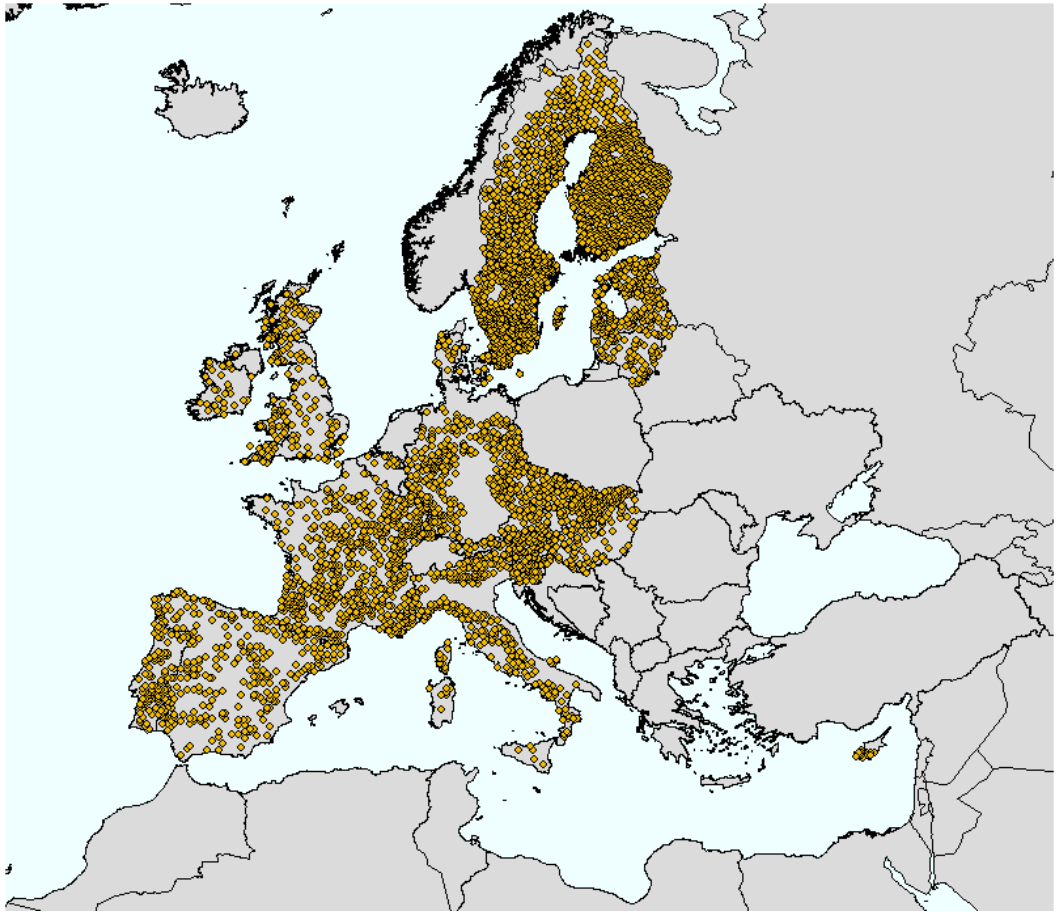


LUCAS data

- LUCAS data is probably the most useful source of information available which can be used to harmonize national datasets
- The eventual presence of a bias between different datasets can be easily detected and corrected using LUCAS data as a reference frame
- The main drawback of the LUCAS data is that is limited to the topsoil and is not covering areas above 1000m of elevation

BIOSOIL data

- BioSoil comprises about 4000 field observations including soil profiles
- Its geographical range is smaller than the one of LUCAS



BIOSOIL data

- BioSoil was undertaken as part of an Administrative Arrangement of the European Commission JRC and Directorate General Environment (DG ENV)
- The aim of the BioSoil project is to demonstrate how a large-scale European study can provide harmonized soil and biodiversity data and contribute to research and forest related policies
- Survey included the description of soil profiles and the sampling of soil horizons
- The main drawback of the BioSoil data is that is limited to forest areas

SPADE/M and Hypres

At European level, two databases with soil profile data exist : SPADE-1 and SPADE-2. The SPADE/M (Soil Profiles) provides an integrated Data structure for 560 profiles (not nearly enough for DSM though)

HYPRES is the database of Hydraulic properties of European Soils contains detailed measured properties of soil profiles for a large range of soil types located in 12 European countries

HYPRES contains profiles data about soil texture (sand, silt, and clay) and organic matter content

Differences between SPADE/M and Hypres

In SPADE/M a soil profile is characterized by a continuous and non-overlapping vertical sequence of horizons. The profile data may be the result of the analysis of several profiles or horizons, but in this case only the aggregated data are used to characterize the profile and stored in the database

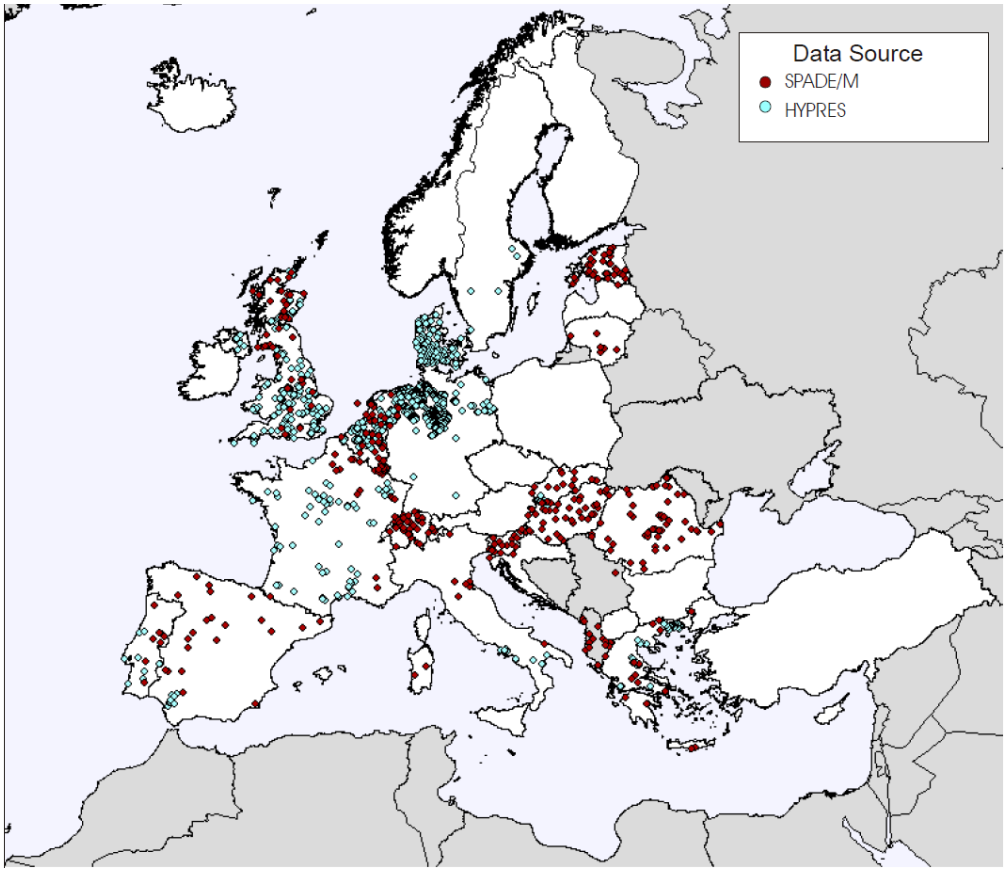
In contrast, in the HYPRES database a soil profile can be characterized by duplicate, overlapping or a non-continuous sequence of horizons (don't ask me why…)

Harmonized SPADE - HYPRES

Data from SPADE and HYPRES has been harmonized by Roland Hiederer at Jrc

However the number of profiles which could be "matched" is relatively low: 817

Moreover the geographical distribution of profiles is heavily clustered



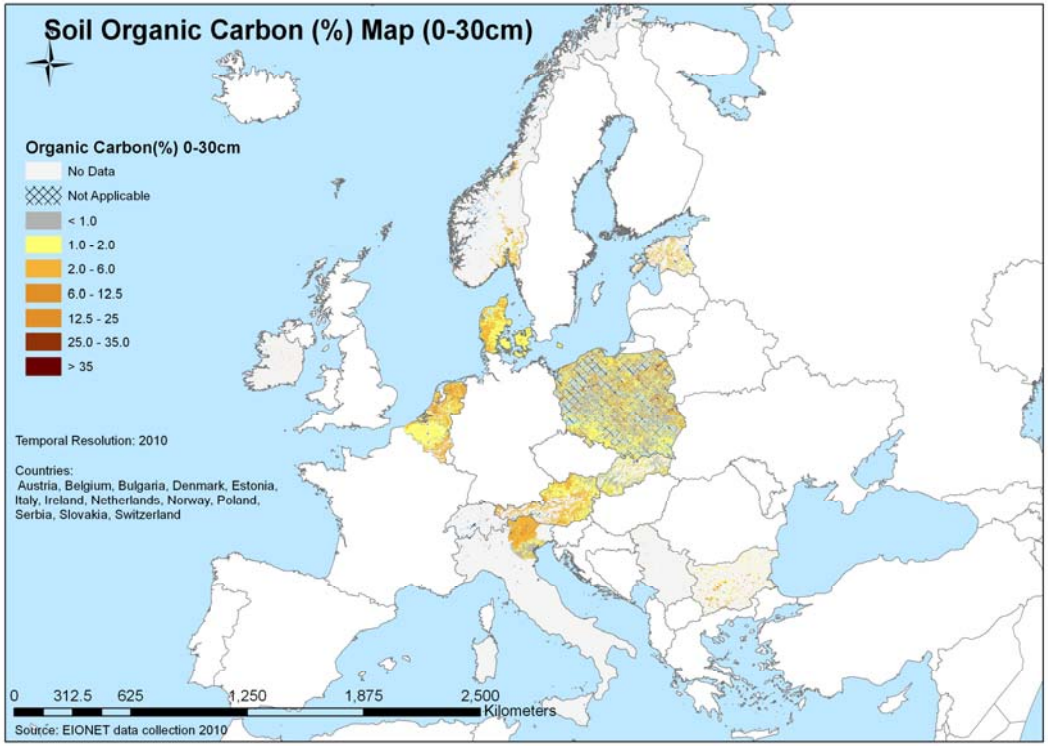
EIONET data collection

European Environment Information and Observation NETwork

Collection and organization of data, development and dissemination of information concerning Organic Carbon and Erosion in Europe

When asked 8 countries provided datasets (Austria, Belgium, Bulgaria, Denmark, Italy, Netherlands, Poland, Slovakia) and 4 countries only points (Estonia, Norway, Serbia, Switzerland)

The EIONET project shows that getting soil data from different countries and regions might be difficult



A summary about point data

The available point data coming from measured samples consists of:

- 20000 topsoil measures from LUCAS project with a full range of chemical and physical measures
- 4000 profiles from BIOSOIL with a full range of chemical and physical measures
- ~1300 profiles from SPADE/HYPRES with data about soil texture and soil organic matter

Part two: the work being done…

At present most of the focus has been on:

1. Harmonizing soil data
2. Estimating extrapolation error over large distances

1) Some ideas… harmonization

The main issue about using local datasets or data from different surveys is the possible bias among datasets

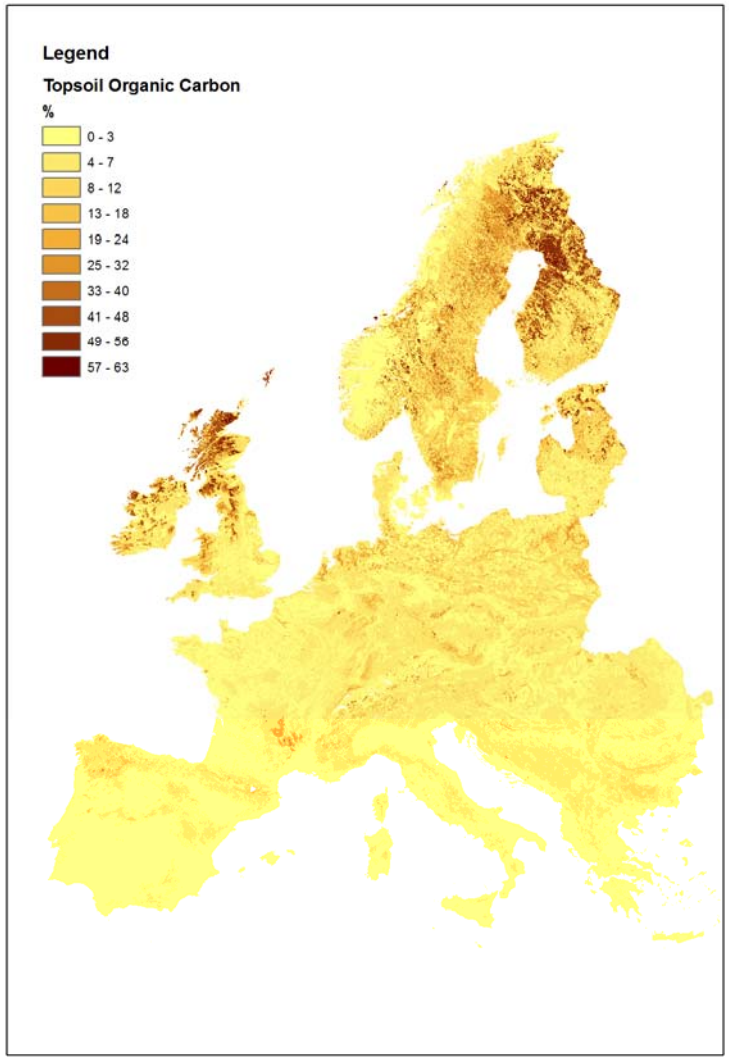
EU wide harmonized measures alone might not be sufficient to produce high resolution maps, however can be used as a reference measure to correct bias in local datasets

Example: correcting Soil Organic Carbon Estimates derived from the European Soil Database

1) Some ideas…

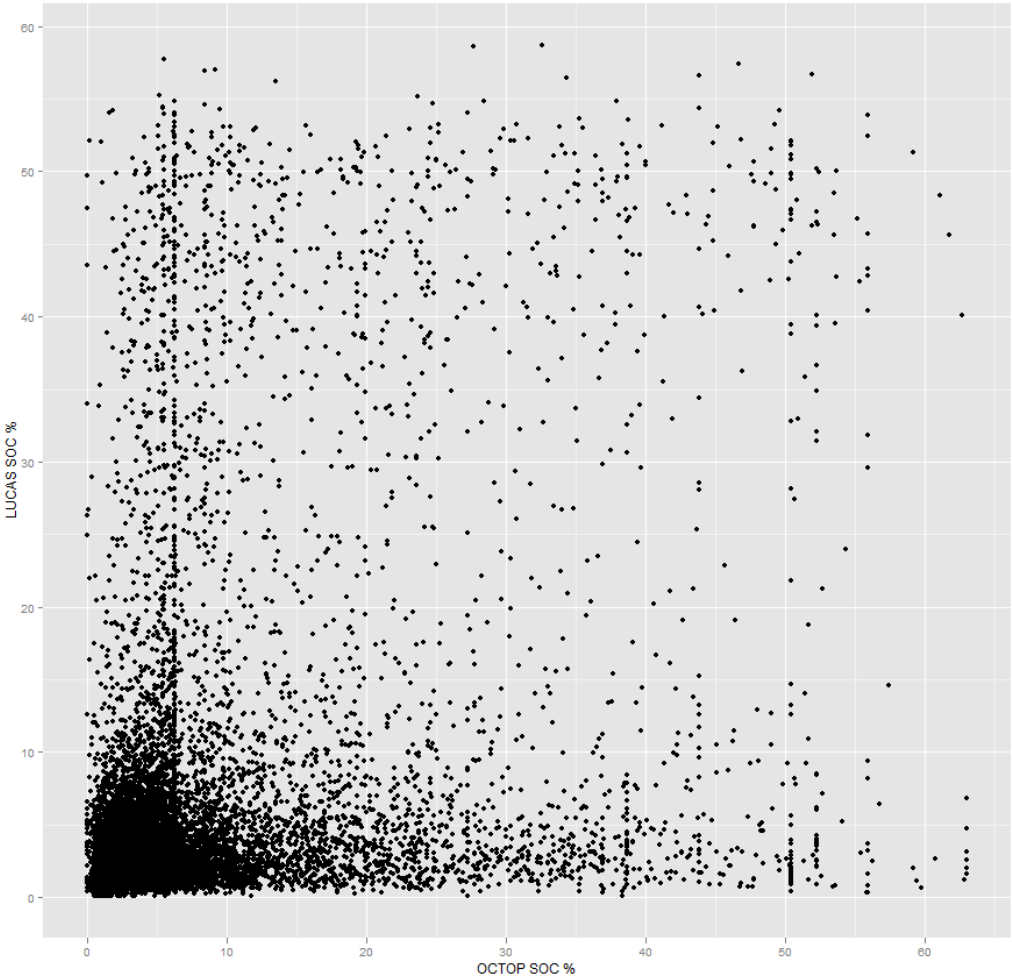
The Topsoil Organic Carbon content map was derived from the European Soil database

The LUCAS data can be used to test if the prediction have the same reliability all over Europe



1) Some ideas…

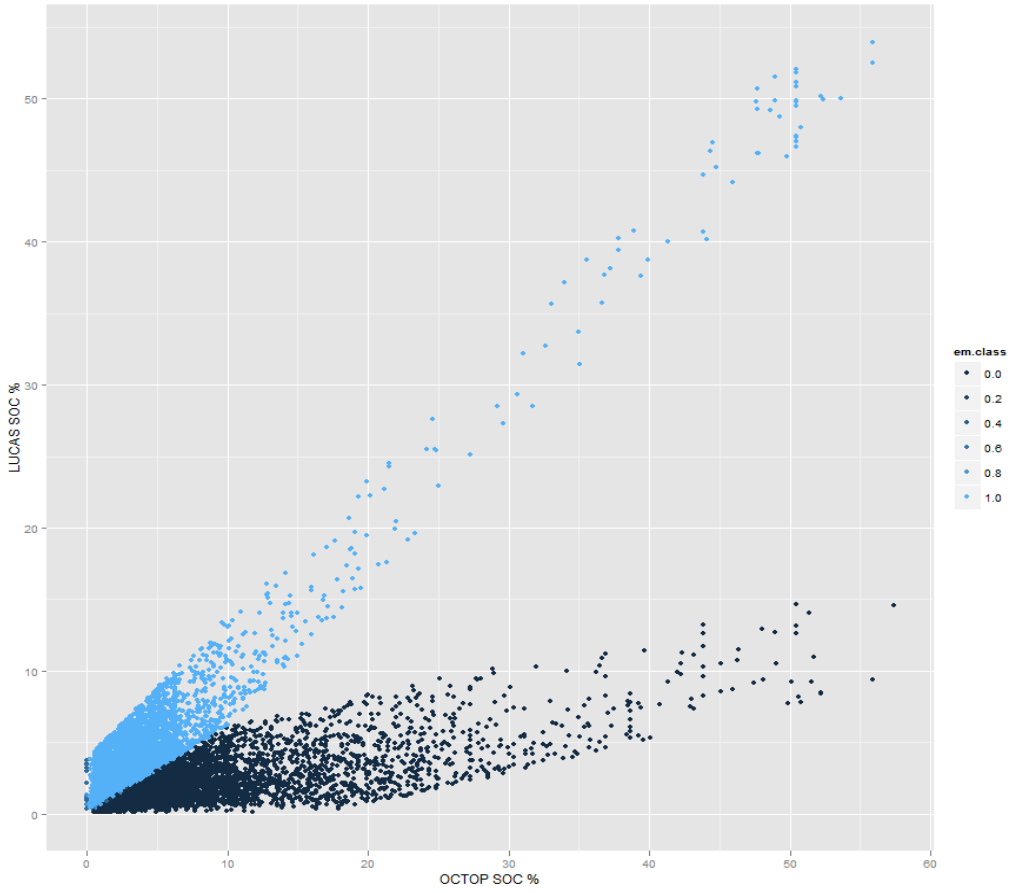
- LUCAS and ESDB OC values are seemingly unrelated
- However, looking at the data, some “trend” appears to be present
- It would be interesting to evidence if these different trends are spatially clustered…



1) Using the EM algorithm to identify clusters of mismatched data

The Expectation Maximization algorithm can be used to fit a mixture of linear regressions

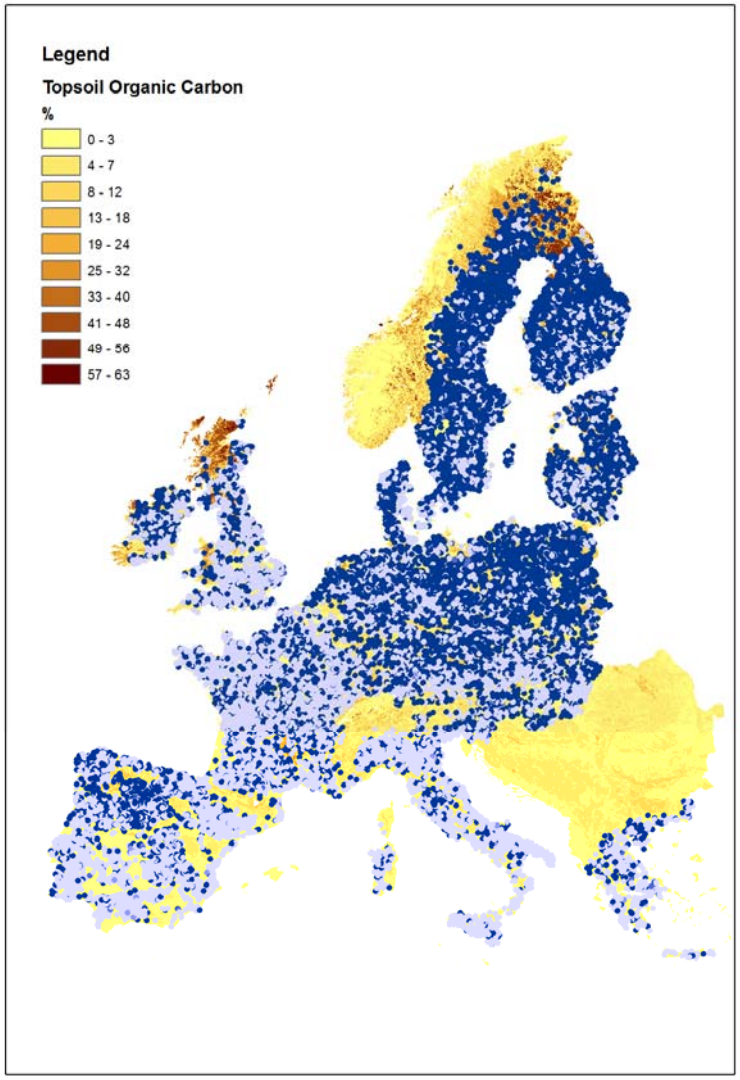
It separates different regression models within the data (still working with a mixture of two…)



1) Spatial distribution of EM clusters

- Plotting the data from the two clusters shows that the presence of different trends seems to arise from differences in the original national datasets

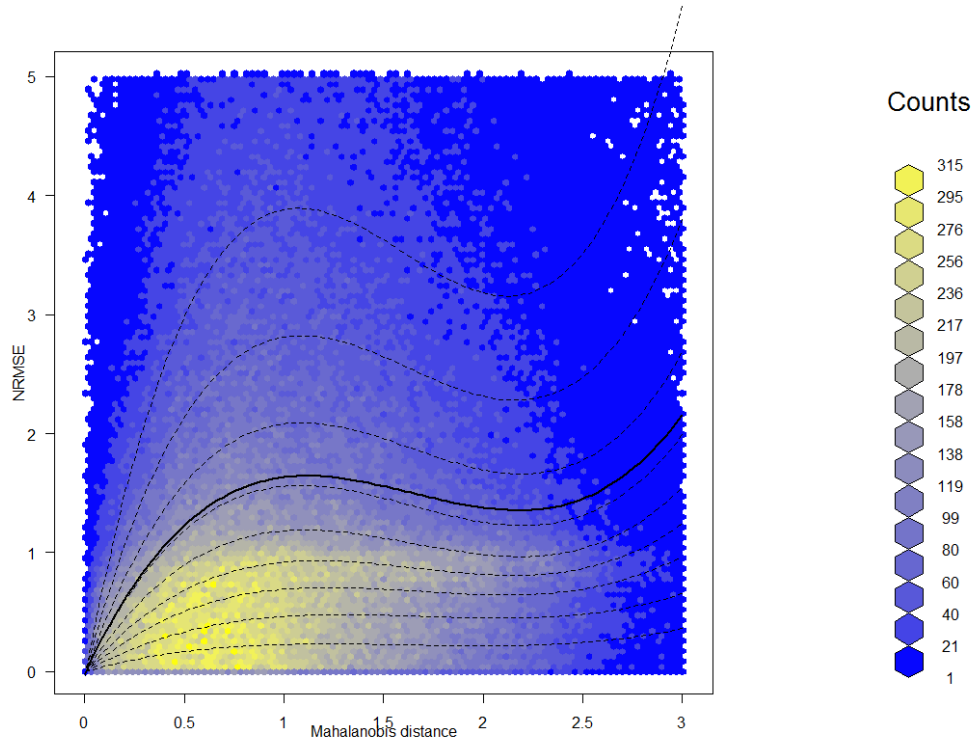
- Northern and Eastern Europe SOC is probably overestimated and should be corrected



2) An idea about error estimation

Still a work in progress…

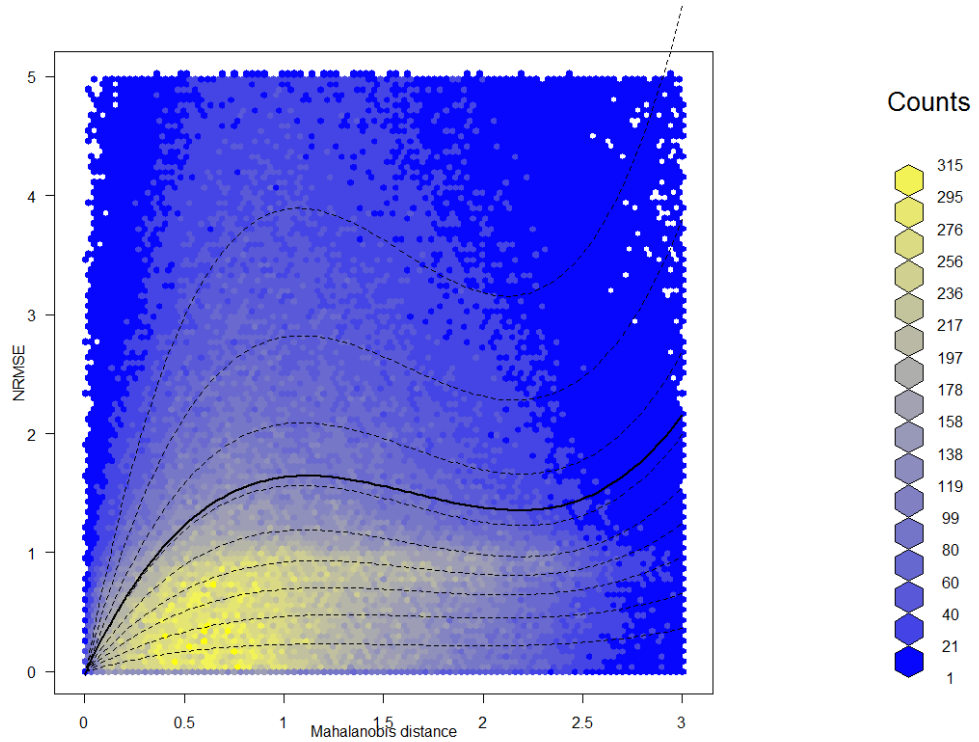
The idea is to assess the error in predicting a given soil property as a function of the Mahalanobis distance between the data used to fit the model and the points/pixels where data is predicted



2) An idea about error estimation

The procedure is a sort of bootstrapping:

1. *Take a random centroid in feature space*
2. *Find the n closest points whose distance is below a given threshold*
3. *Use those points to fit a weighted regression model*
4. *Apply the model to predict k randomly selected points*
5. *Estimate the error and relate it with the Mahalanobis distance between n and k points*
6. *Repeat m times*

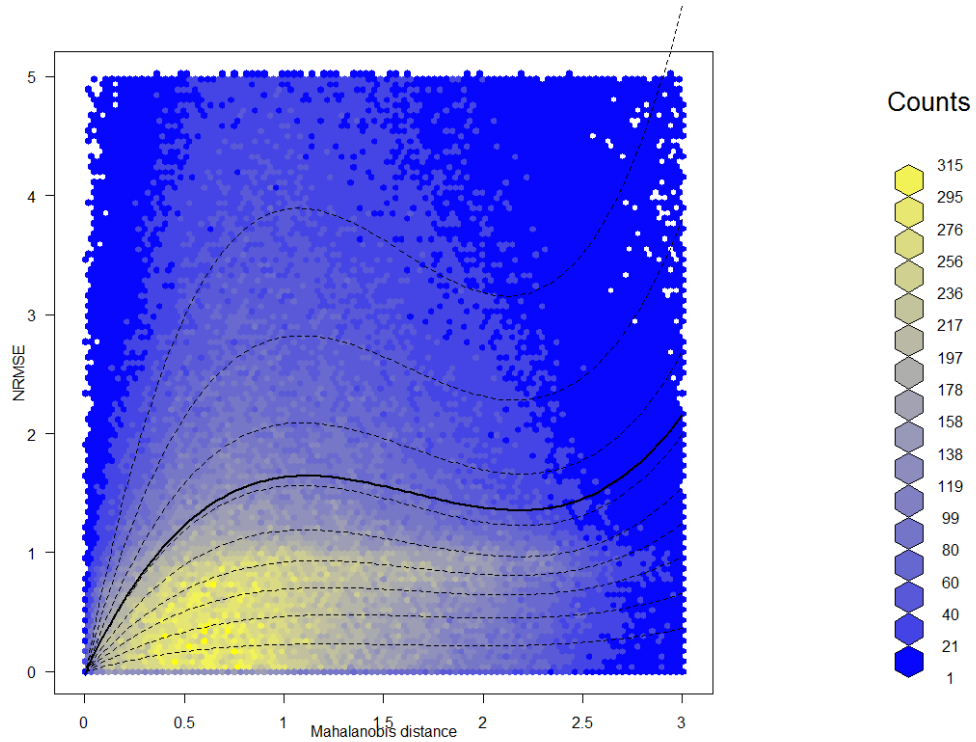


2) An idea about error estimation

Mapping the error in geographical space will require to assign the error estimate to all the pixels in the map on the basis of their Mahalanobis distance

The procedure is iterative, basically a sort of sampling from the error/distance distribution

Should avoid using global error estimates over large areas



Final remarks… about the data

The presence of EU wide datasets will help the harmonization of national data and provide additional information for DSM in Europe

Nevertheless, at present, a reasonable amount of soil profiles data is missing; this data must come from national inventories

Existing soil maps, while potentially useful, might be better used as prior information (in the Bayesian sense) and not as discrete covariates in the mapping process

Final remarks… about error estimation

At the moment the activity at Jrc is mostly focused on data harmonization

Given the nature of the available data, at the moment correcting the bias is probably more critical than estimating uncertainty

As soon as more data becomes available/usable we will focus more on uncertainty estimation (suggestions are welcome)

Thanks for your attention