# Knowledge discovery from soil samples using the partial dependence of random forest under fuzzy logic

**Canying Zeng, A-Xing Zhu, Lin Yang**

**Nanjing Normal university**

# Outline

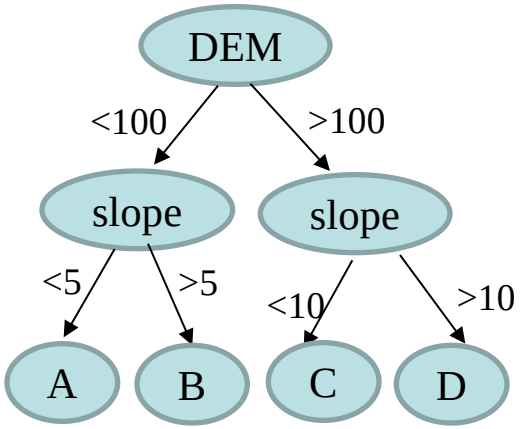
1. Introduction
2. Methodology
3. Case study
4. Conclusions

# 1 Introduction

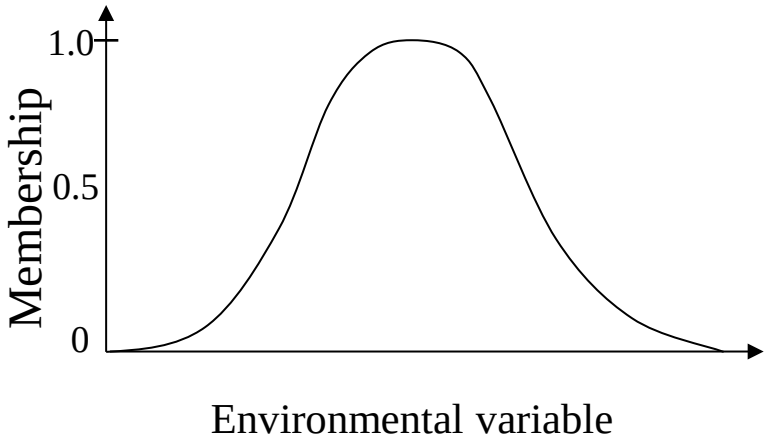
- **The soil-landscape relationship**

**The quantitative knowledge on soil-landscape relationship such as rules or membership function is important for understanding soil, digital soil mapping, and land resource management.**

Rules



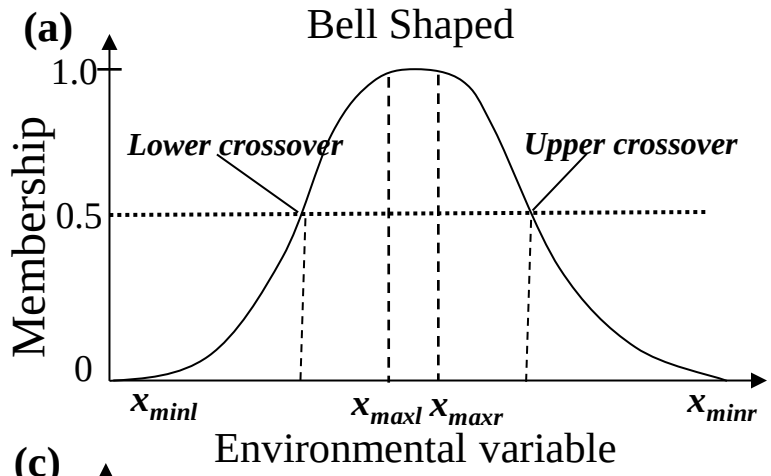

Membership functions

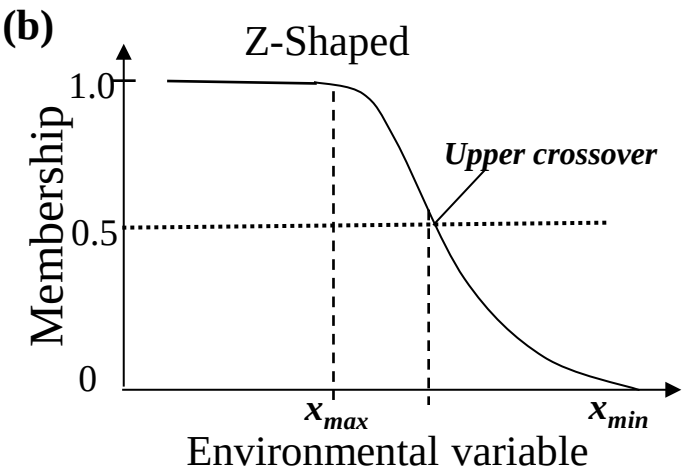


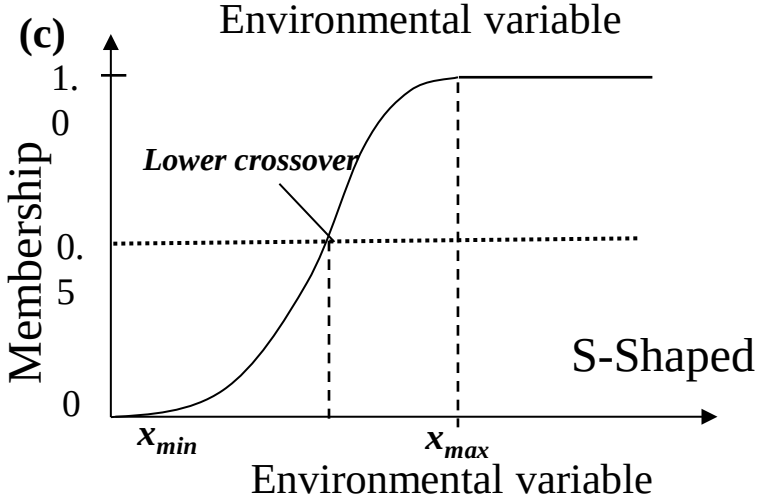
# 2 Introduction

## ◆ Fuzzy membership function

**Membership function is an effective tool to express such knowledge on soil-environment relationships (Zhu 1997, 2001)**




**Three basic form of fuzzy membership functions (Zhu, 1997)**

# 1 Introduction

## ◆ Fuzzy membership function

**Soil samples imply the knowledge of relationships between soils and their underlying environmental conditions**

Random forest

Black-box method, widely used in DSM. It doesn't produce explicitly knowledge on soil-environment relationships.

Partial dependence (*Pd*) plot

It can produce the *Pd* plot, which implies the relationships between soil and environmental.

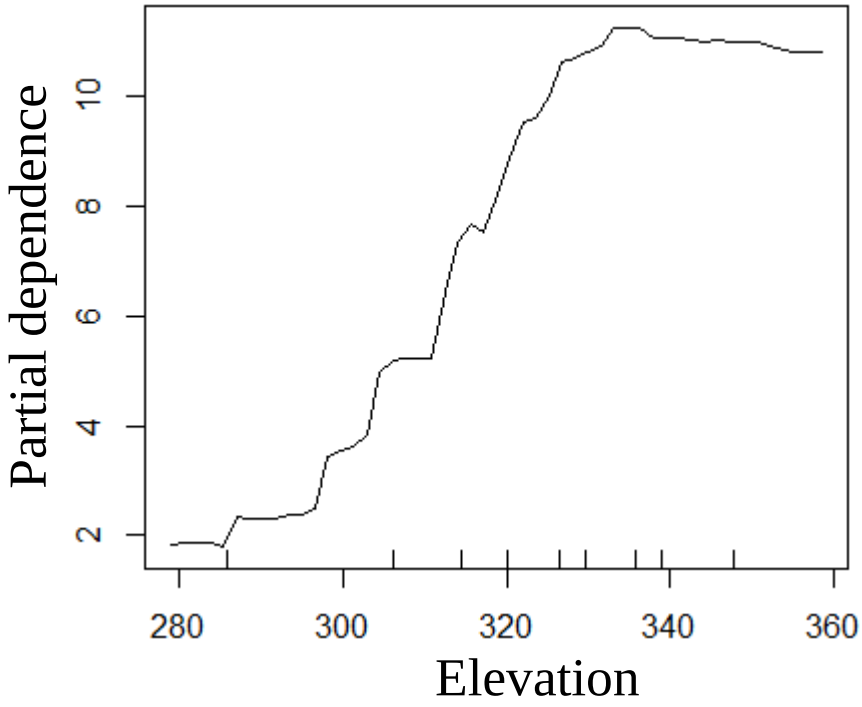
**Partial dependence (*Pd*) gives a quantitative depiction of the dependence of an environmental variable on the class probability (Friedman, 2001)**

# 1 Introduction

- **Random forest and partial dependence**

Species
Land use
…

Environmental variables



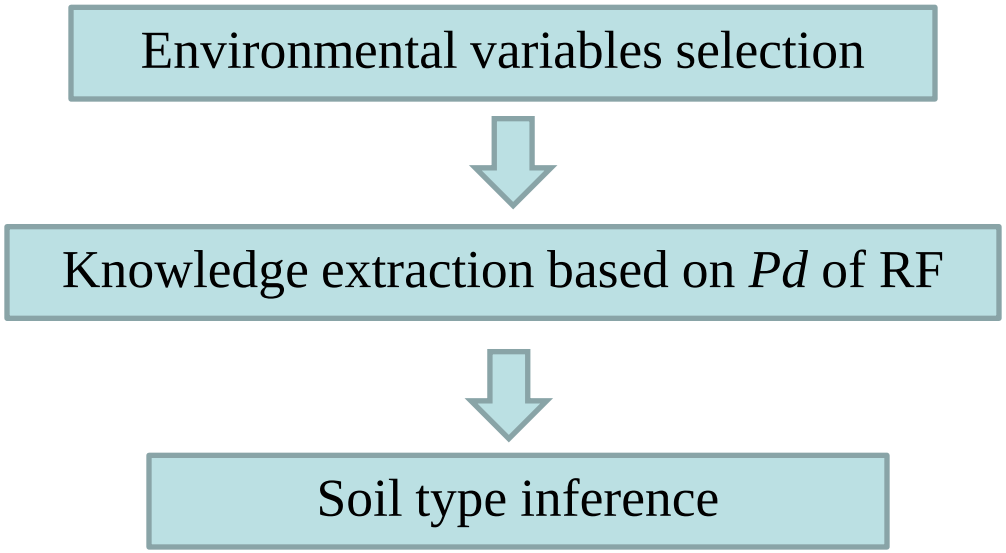

**Some studies used *Pd* plots to explain the relationships between species or land use and environmental variables (Wang et al., 2016; Cao et al., 2015; Cutler et al., 2007), but it cannot be directly used for mapping.**

# 1 Introduction

- **Question**

**How to translate partial dependence plots into explicit membership functions and use it for soil mapping?**

# 2 Methodology



# 2 Methodology

## ◆ Environmental variables selection

**The mean decrease accuracy was used to choose the relevant environmental variables.**

- **The overall mean decrease accuracy (MDA) of the environmental variable in RF should be larger than 0.**
- **NMDA% should be smaller than 10%.**

| MDA(%) | Soil types | | | Overall MDA (%) | NMDA% (%) |
|---|---|---|---|---|---|
| | A | B | C | | |
| Variable 1 | 23.71 | 20.26 | 31.95 | 28.90 | 0 |
| Variable 2 | 20.58 | -4.21 | -0.86 | 19.30 | 20.27 |
| Variable 3 | 5.23 | -3.02 | -0.89 | -0.03 | --- |

# 2 Methodology

## ◆ Knowledge extraction based on *Pd* generated by RF

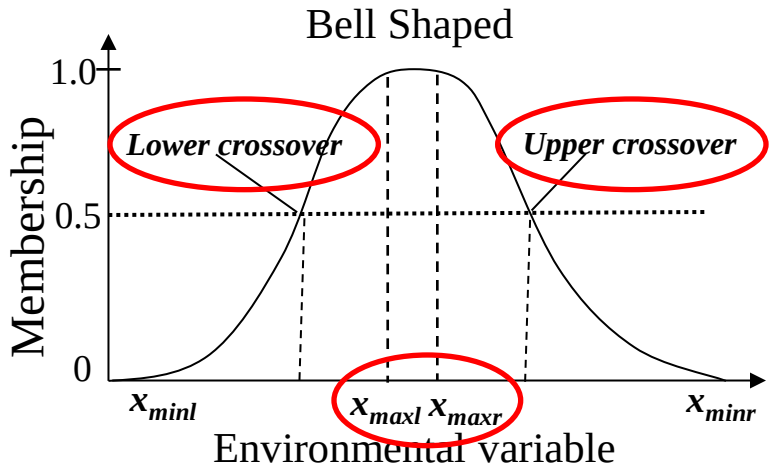
**For each value of an environmental variable, its *Pd* is defined as the proportion of votes for a certain class minus the average proportion of votes for the other classes based on the random forest classifier(Friedman, 2001) .**

The stronger the partial dependence of a value for some variable, the higher probability of the soil existing in this value of this variable.

Partial dependence ⇒ Membership function

↓

Determine the key parameters for membership function

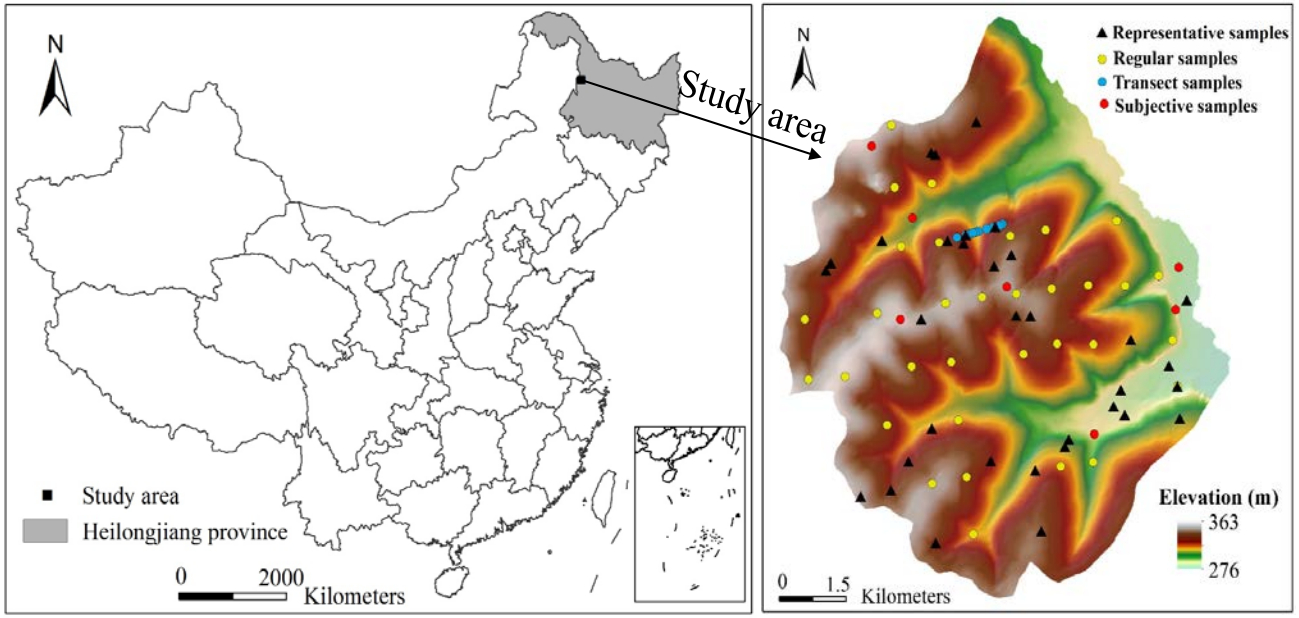



After all the fuzzy memberships between each soil type and environmental variables were constructed, The SoLIM was used to predict soil types.

# 3 Case study

## ◆ Study area

**The study area is located in Heshan farm of Nenjiang County in Heilongjiang province (60 km$^2$). Its elevation ranges from 276 to 363 m. The land use and soil management is generally uniform across the study area.**




**A: Pachic Stagni-Udic Isohumosols**
**B: Mollic Bori-Udic Cambosols**
**C: Typic Hapli-Udic Isohumosols**
**D: Typic Bori-Udic Cambosols**
**E: Lithic Udi-Orthic Primosols**
**F: Fibric Histic-Typic Haplic Stagnic Gleyosols**

# 3 Case study

## ◆ Environmental Variables

### Description of environmental variables

| Variables | Module | Softwares |
|---|---|---|
| Elevation | Elevation | Acgis 10.1 |
| Slope | Slope in ArcInfo | Acgis 10.1 |
| Cosaspect | Cos(Aspect) | Acgis 10.1 |
| Planc | Plan curvature (Shary et al., 2002) | Acgis 10.1 |
| Profic | Profile curvature (Shary et al., 2002) | Acgis 10.1 |
| TWI | Topographic Wetness Index (Qin et al., 2011) | SimDTA |
| Hand | Height Above the Nearest Drainage (Gharari et al., 2011) | Python |
| TCI | Terrain Characterization Index (Park and van De Giesen, 2004) | SimDTA |
| TRI | Terrain Ruggedness Index (S.J. et al., 1999) | SimDTA |
| TPI | Topographic Position Index (Jenness, 2006; Weiss, 2001) | SimDTA |
| Relief | Topographic relief (Skidmore, 1990) | SimDTA |
| Slopepos | Fuzzy slope position including Ridge, Shoulder, Back slope, Foot, Channel with value of 1-5 (Qin, Zhu, et al., 2009) | SimDTA |

We generated twelve environmental variables commonly used in this study area.

# 3 Case study

## Evaluation

**Two scenarios were conducted to test the effectiveness of proposed method with different training samples.**

### Scenario 1

**Training: 33 representative samples**
**Validation: 50 samples as validations**

Compared with the knowledge extraction method of Yang et al. (2013) base on same training samples and validation data.

Randomly split

### Scenario 2

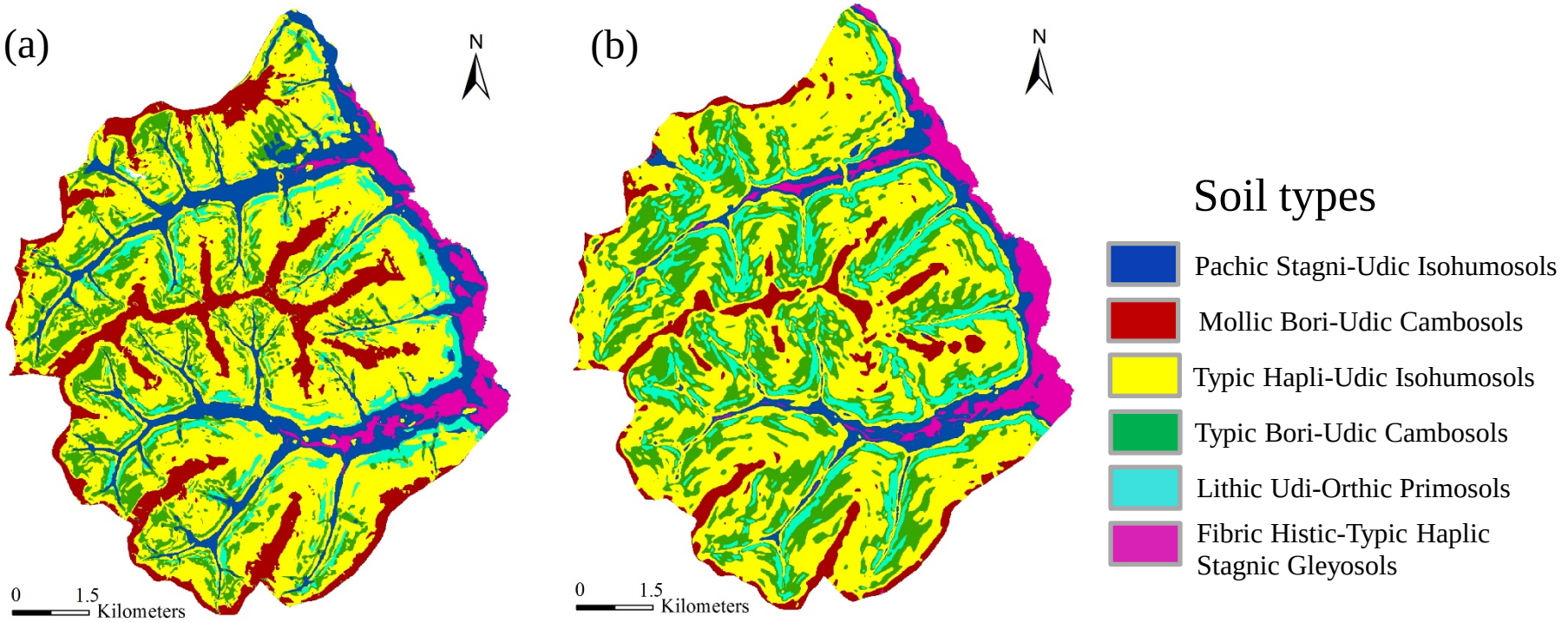
**Training: 2/3 samples**
**Validation: 1/3 samples**

The split was conducted nine times and thus nine sample sets were generated.

**Random forest was also conducted based on each training samples for the two scenarios.**

# 3 Case study

## ◆ Results of Scenario 1




**Mapping results using (a) membership function and (b) random forest**

## Prediction accuracy

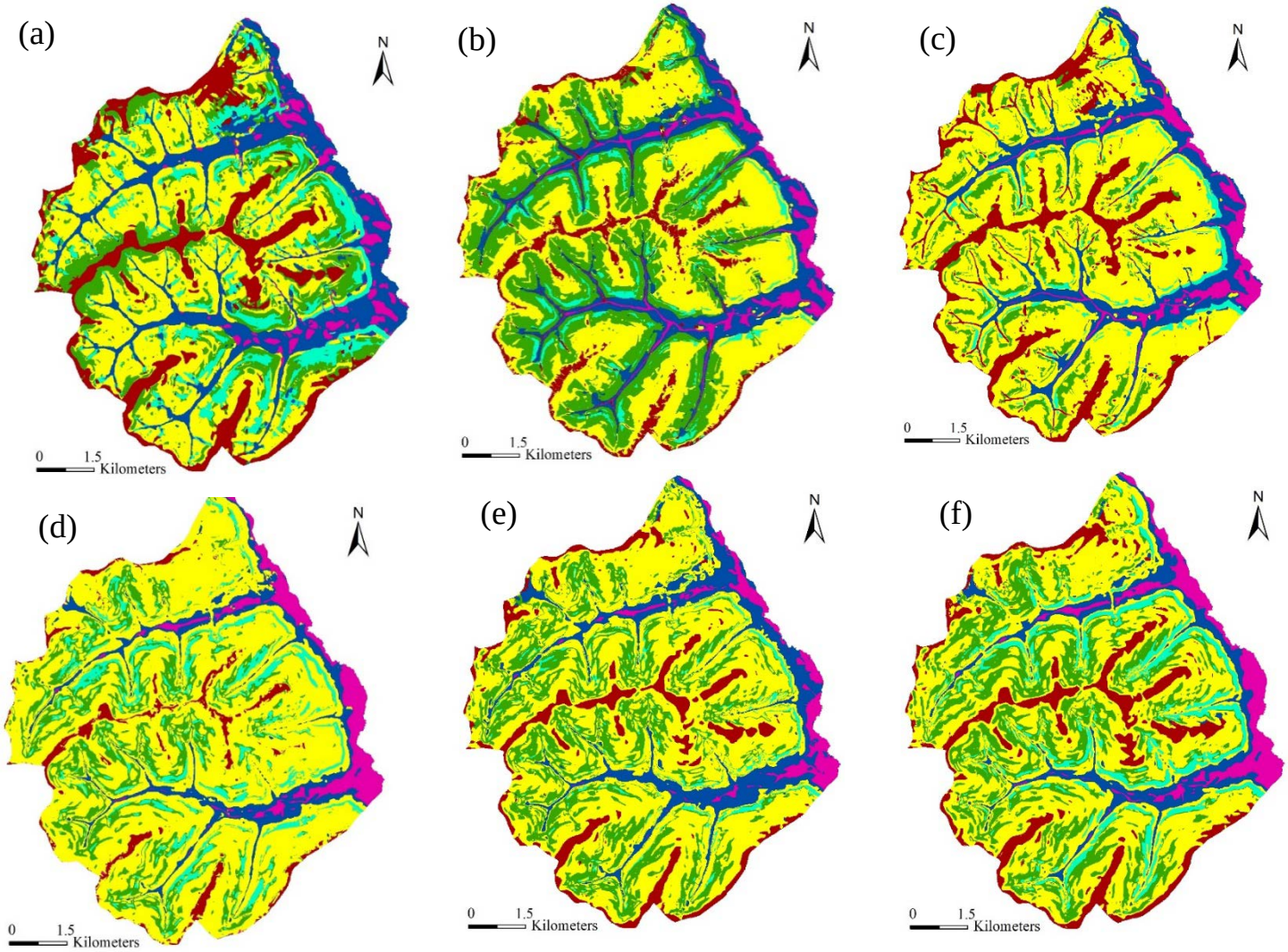
**Using membership function: 0.78**

**Using random forest: 0.60**

**Yang et al. (2013): 0.76**

# 3 Case study

## ◆ Results of Scenario 2

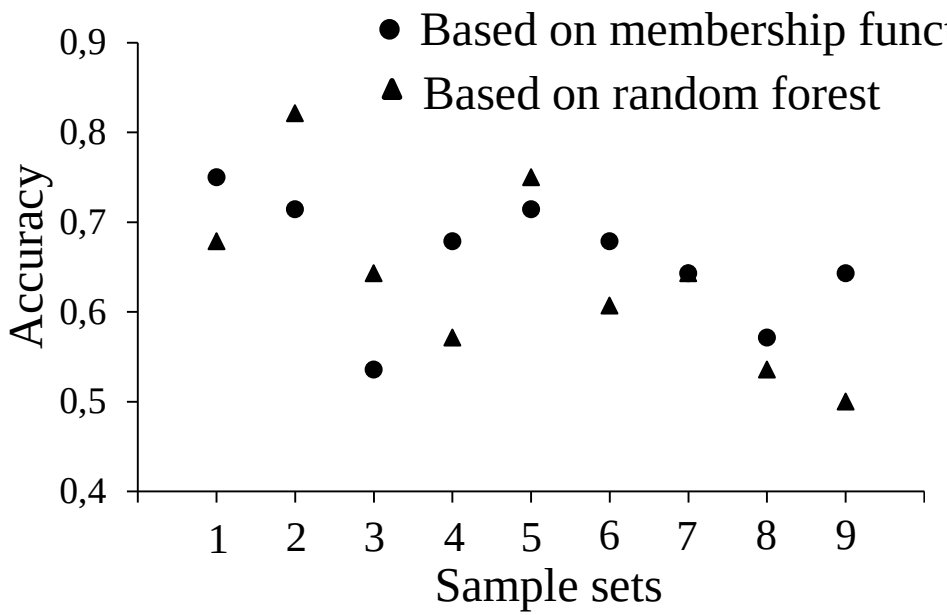



Sample set 1 Sample set 2 Sample set 3

**Mapping results using (a)-(c) membership function and (d)-(f) random forest**

# 3 Case study

## Results of Scenario 2




| | Average accuracy | STD |
|---|---|---|
| ⇒ | 65.87% | 6.92% |
| ⇒ | 63.89% | 9.59% |

**The prediction accuracy for each sample set of scenario 2**

- **The prediction accuracy of scenario 2 based on the memberships are lower than 1. Because adding atypical samples into training samples made the membership curves wider and overlapped larger. Transitional areas possessed high membership to certain soil types.**
- **Sometimes, the accuracies of random forest were high and sometimes were low. The possible reason is that the random forest was more prefer the training samples with the more widely coverage of environmental conditions for each soil type.**

# 4 Conclusions

- **The knowledge of relationships between soil and environmental variables can be extracted from partial dependence of random forest. The extracted knowledge is effective to predict soil types in the study area.**

- **Training samples will greatly impact mapping results and accuracies.**

  **Using representative samples as training samples is recommended when applying the proposed method to extract soil-environment knowledge.**

- **Training samples with a full coverage of environmental conditions where each soil type distributes would benefit random forest to obtain more accurate soil maps.**

***Thank You!***