# Combining heterogeneous across country data on proxies for prediction of methane in dairy cattle

*Enyew Negussie*

*O. González Recio,   F. Biscarini*

**Y. de Haas, N. Gengler, H. Soyeurt, N. Peiren, M. Pszczola, P. Garnsworthy, M. Battagin, A. Bayat, J. Lassen, T. Yan, T. Boland, B. Kuhla, T. Strabel, A. Schwarm, A. Vanlierde**

**@METHAGENE**

Luke
NATURAL RESOURCES INSTITUTE FINLAND

# Introduction

- **Methane ($CH_4$) in ruminants**
  - Product of normal digestion
  - Ruminants **- lose 2-12% GE**
  - One most **potent GHG**
  - Mitigation - nutritional and environmental benefits

**Basis of methane emission ?**

**•Nutrition & Genetics**

- **Success in mitigation depends =** ability to measure trait
- **$CH_4$ measurement difficult, costly**

Costly & logistically difficult

$h^2_{Sheep}$=0.1-0.3

$h^2_{Beef}$=0.2-0.4

$h^2_{Dairy}$=0.1-0.2

# Prediction models - as alternative

- **Most prediction models – developed so far:**
  - Most models valid = only to their circumstances
  - Data of treatment means from different studies
    - Different treatment means = varying uncertainties

- **Developing robust prediction models is essential**
  - direct individual animal measurements
  - from across - countries & heterogeneous sources
  - applying – machine learning techniques are lacking

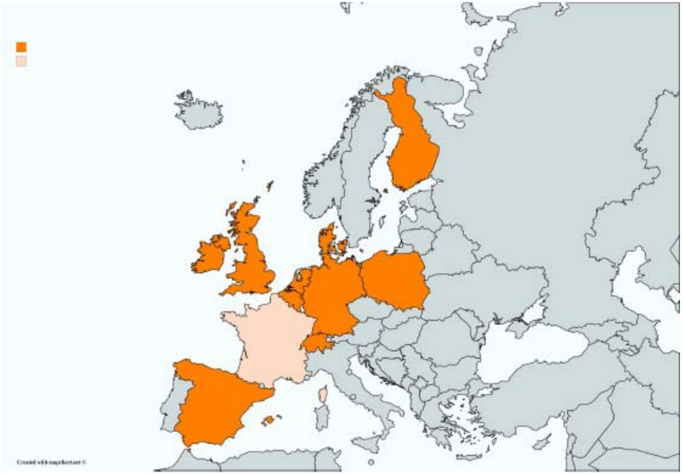
## Objective

- **Combine heterogeneous individual animal proxies on $CH_4$**
- **Develop robust $CH_4$ prediction model**

# Data

- ➢ **Across 11 European countries**
  - ➢ Different production systems
  - ➢ 16 different data sets, 3000 animals, 65 000 obs.

- ➢ **Direct animal measurements**
  - ➢ Methane & **proxies** for methane
  - ➢ Diverse methods, breeds, parity, age groups etc.



**Contributions**

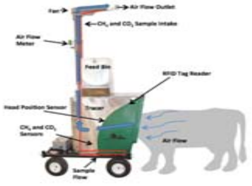
1. Germany
2. Finland
3. Netherlands
4. Belgium
5. Poland
6. Switzerland
7. UK
8. Ireland
9. Spain
10. Denmark
11. France

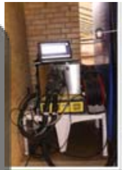
# Data composition - heterogeneous

- **Species/Breeds**
  - Data mainly dairy, beef, sheep
  - HOL, Nordic Red, Norwegian Red & Crosses, Brown Swiss

- **Parity & stage of lactation**
  - Parity 1 – 3+,   DIM 5 - 365

- **$CH_4$ measurement methods**
  - **Cattle respiration chamber**
  - $SF_6$
  - GreenFeed
  - **Sniffers:**
    - NG-guardian, Gasmet, F10, etc.





# DATA ANALYSIS

# Step1. Harmonization & standardization

- Data standardization & normalization (coding, trait definition, units etc.)
- **Target – $CH_4$ (g/d)**
- Core traits/proxy/predictors common to most datasets

**Core proxies/variables**

DMI, kg/d
Milk, kg/d
Prot, %
Fat, %
Recording date
$CH_4$ method

Herd
Parity, no.
BWT, kg
Calving date
Breed, Age
DIM, d

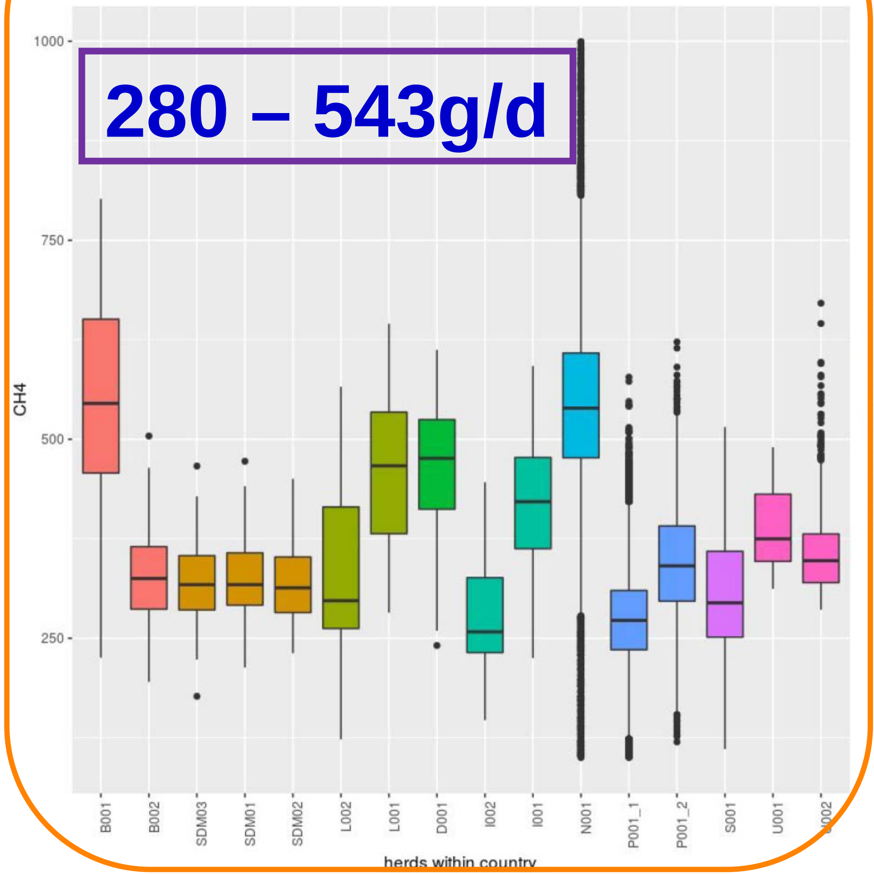
**Final data**

**~50 000 obs.**
**2400 cows, 9 countries, 16 herds,**
**6 breeds,**
**4 $CH_4$ methods,**
**1-3+ parties, dim 5-365**

## Data descriptive stat

| Variable | # obs. | Mean |
|---|---|---|
| Dry matter intake, kg/d | 912 | 17.9 |
| Milk Yield (kg/d) | 48802 | 33.5 |
| Milk Protein % | 38091 | 3.3 |
| Milk Fat % | 38125 | 3.8 |
| Body weight (BW) kg | 48641 | 571.0 |
| Methane ($CH_4$) g/day | 48804 | 347.0 |

## $CH_4$ across-datasets


**280 – 543g/d**

# Step 2. Data Analysis : Machine learning

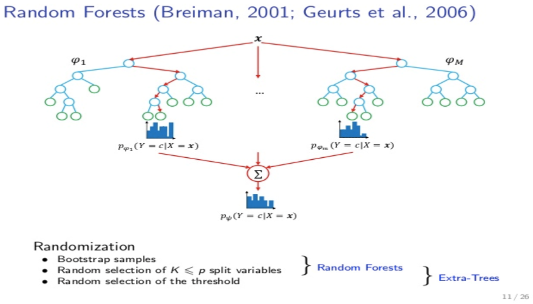
**Principal comp. analysis (PCA)**

- Exploratory, visualization of high dimensional data

**Random Forest (RF)** - Machine learning algorithm

- Decision trees on bootstrapped samples of data
- Averaging each estimate to make final prediction
- Able to capture complex interaction structures in data
- Robust to over-fitting & reduce error of prediction
- Provide – relative importance of each proxy/variable
- Reg. Trees: herd, breed, Lact. status, milk, fat, prot, dim, CH4 method, BWT, DMI

- **Prediction accuracy: 10-fold Cross validation**
  - With-in herd – DIM vs no DIM
  - Between herd



Random Forests (Breiman, 2001; Geurts et al., 2006)

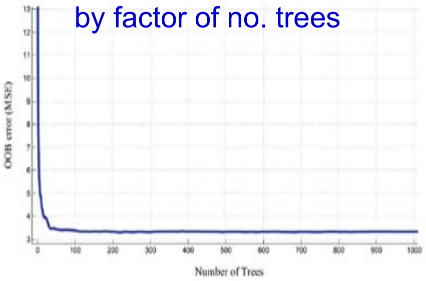
Randomization

- Bootstrap samples
- Random selection of $K \leqslant p$ split variables
- Random selection of the threshold

} Random Forests

} Extra-Trees

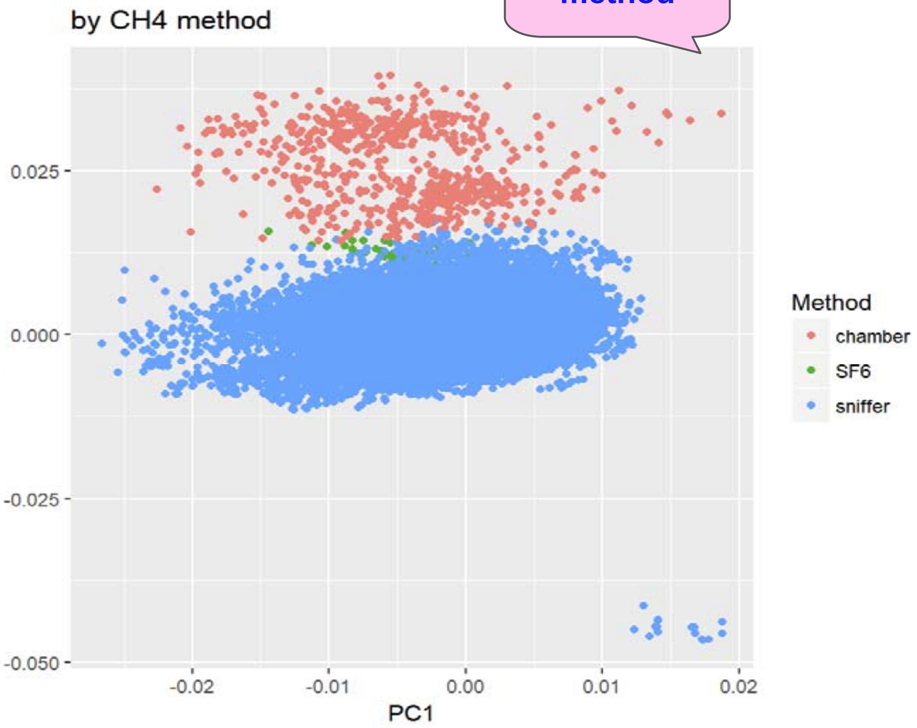
11 / 26



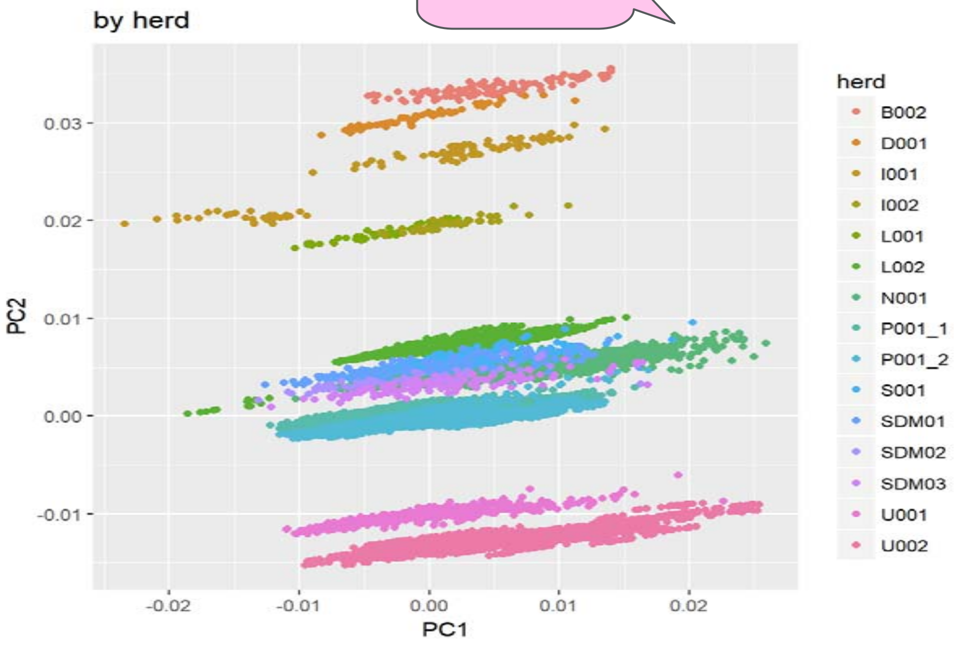
Reduces prediction error by factor of no. trees

# Results - PCA

CH4 method



Herd



# Prediction accuracy

| Herds/ $CH_4$ method | Within-herd accuracy (no DMI) |
|---|---|
| Chamber | 0.57 – 0.80 |
| Sniffers | 0.26 – 0.82 |

For some sniffers within-herd accuracy was as high as for chambers

| Herds/ $CH_4$ method | Within-herd accuracy | | |
|---|---|---|---|
| | no DMI | | With DMI |
| Chamber | 0.53 | 38% ↑ | 0.86 |
| Sniffers | 0.27 | 36% ↑ | 0.42 |

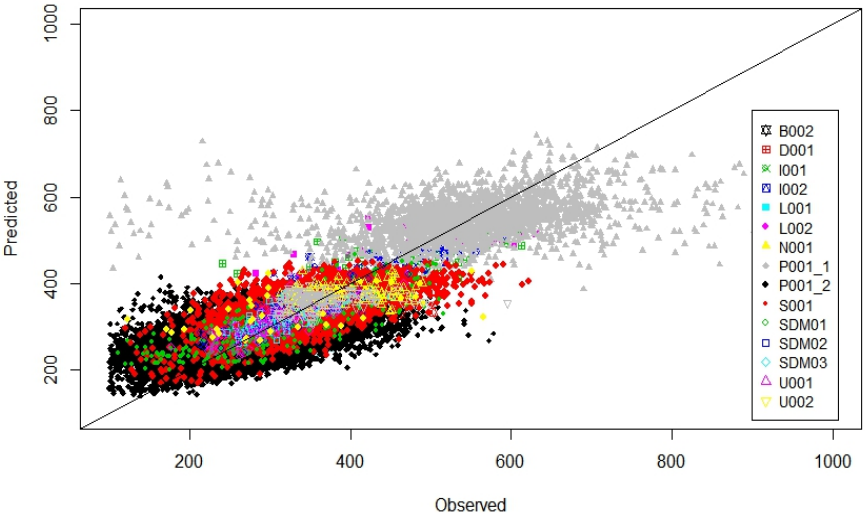
DMI markedly improved within-herd accuracy of both – more DMI needed

# Prediction accuracy (within vs Between herd)

| Herds/ $CH_4$ method | Within herd | Between herd |
|---|---|---|
| **Chamber** | 0.59-0.87 | 0.43-0.80 |
| **Sniffers** | 0.26-0.82 | 0.10-0.55 |

Each herd/method has its own peculiarities – more harmonization

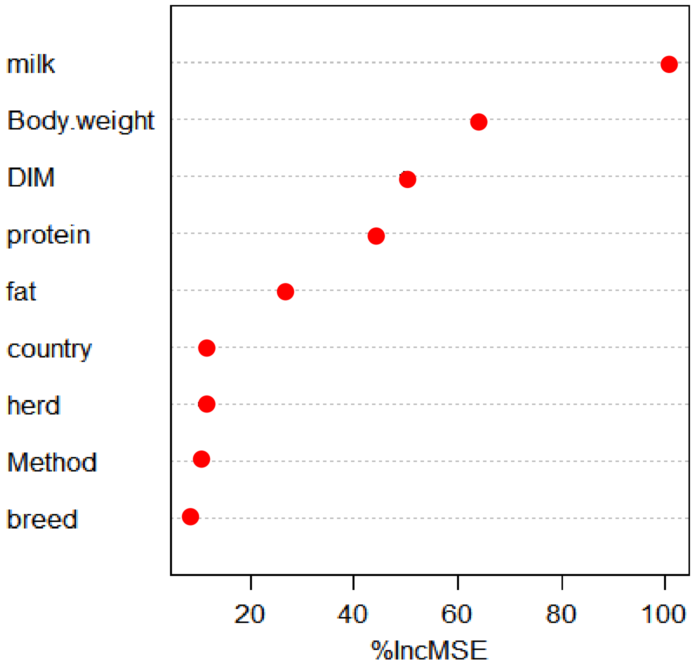
# Accuracy - Overall



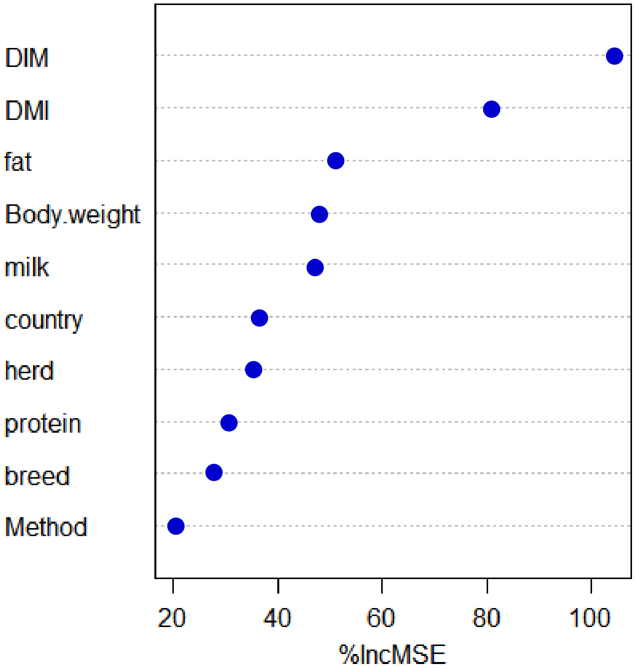
Overall accuracy 0.813 *Vs* 0.814 when DMI included
Caution!

# Variable importance

## No Dry matter intake



## Dry matter intake included



**DMI & production traits at the top whilst, breed and $CH_4$ methods at bottom of the regression tree**

# Conclusions

- **Combined proxies** **- Accuracy without DMI 0.81**
  - well recorded proxies have shown great potential
- **Variable importance**
  - DMI, BW and milk traits (fat & protein) - most important proxies
  - Breed & $CH_4$ measurement method at the bottom tree
- **Methods & techniques**
  - future use combining pedigree & genotype data for across-country joint genomic predictions for envt'al impact traits

- **Luke – GreenDairy project**
- **METHAGENE COST Action FA1302 of the EU**
- **METHAGENE WG3 working group**

# Thank you!

NATURAL RESOURCES
INSTITUTE FINLAND