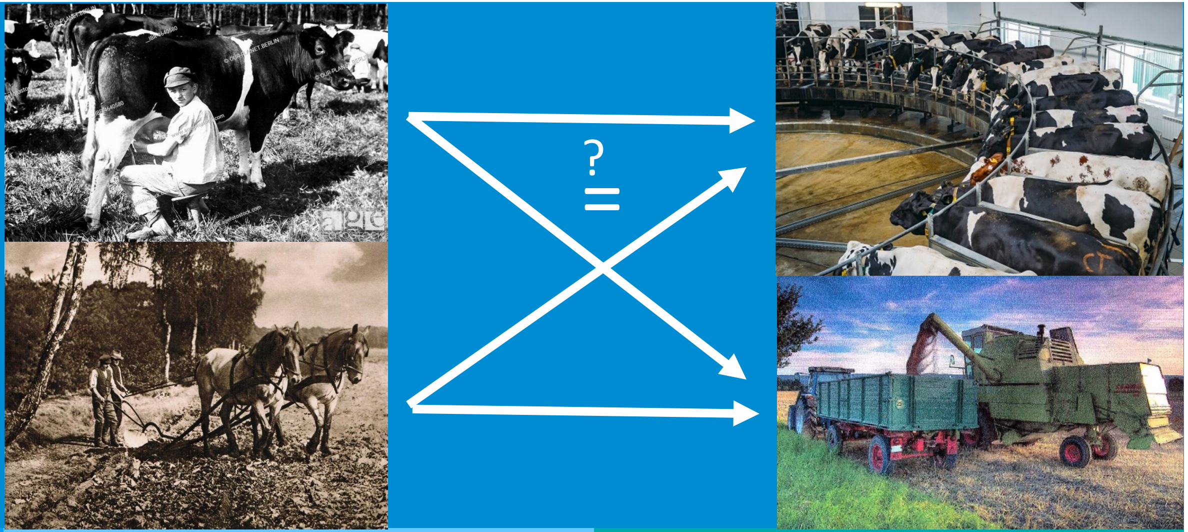
Long time series of identical farms in FADN - Can machine-learning help to consolidate farm identifiers?

Frank Offermann and Heiko Hansen
Thünen Institute of Farm Economics



Pacioli Workshop 2023
Ptuj

Introduction

Longitudinal/panel data (observations identified over time) is key data for many agri-economic studies

- ➢ 300 paper in Ag-Econ journals in last five years
- Identifier often based on ‚farm‘ as central concept in agriculture – and its data bases

Major challenges

Is this still the same farm?

Ancient problem:

Temporal identity of an immaterial object

(-> Plutarch, Paradox of Theseus Ship)

Temporal identity - Is this still the same farm?

To what extent does the identity of farm require the identity of its parts and/or the identity of its attributes?

- Is a farm still the same farm when it changes some of its characteristics?
- Is a farm which quits some major enterprise, e.g. dairy production, still the same farm?
- Should we keep the farm identifier constant if a farmer retires and his/her offspring inherits the farm? Is the answer the same if the farmer sells the farm to an unrelated person?
- What if a large farm takes over all the assets and obligations of a small farm – clearly, the identity of the small farm is dissolved, but will we think of the large farm as being the same farm as before? (What if the small farm takes over the large farm?)

Introduction

Longitudinal/panel data (observations identified over time) is key data for many agri-economic studies

- ➢ 300 paper in Ag-Econ journals in last five years
- Identifier often based on ‚farm' as central concept in agriculture – and its data bases

Major challenges

Is this still the same farm?

Ancient problem:

Temporal identity of an immaterial object

(-> Plutarch, Paradox of Theseus Ship)

Many studies use secondary data (e.g. FADN)

- Little or no control over the identifier
- FADN has few guidelines on what constitutes an 'identical' farm

FADN identifier

EU FADN

- If there is subdivision, merger or any other fundamental change in a holding, it should be considered as a new holding and assigned a new number.
    - A change in type of farming is not considered enough for assigning a new number.
    - If the [administrative] regional boundaries change, new holding numbers should be assigned

German national FADN

- A new farm identifier should be allocated when a farm changes its legal form.

Introduction

Longitudinal/panel data (observations identified over time) is key data for many agri-economic studies

➢ 300 paper in Ag-Econ journals in last five years

- Identifier often based on ‚farm' as central concept in agriculture – and its data bases

Major challenges

Is this still the same farm?

Ancient problem:

Temporal identity of an immaterial object

(-> Plutarch, Paradox of Theseus Ship)

Many studies use secondary data (e.g. FADN)

- Little or no control over the identifier
- FADN has few guidelines on what constitutes an 'identical' farm

Sample characteristics?

Objective

Objective: Establish consolidated panel data using a ‘broad’ definition of time-invariant identity to systematically analyse the impact of different alternative definitions on sample size and key sample characteristics

Case Study: German national FADN, 1995-2019, >250,000 observations

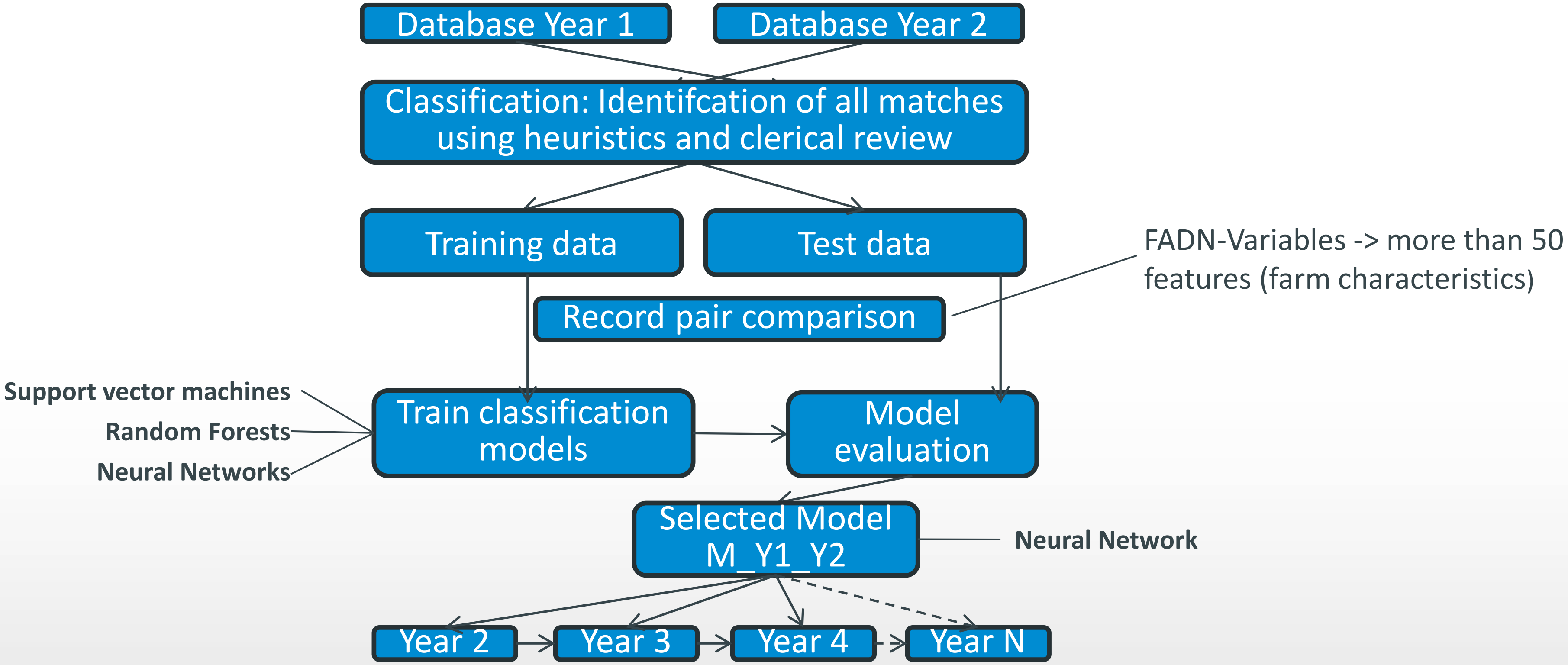
- Poor quality of existing identifier (technical flaws; missing guidelines; heterogenous subjective decisions on identifier)

Challenge: Manual consolidation of identifier tedious and time consuming

Clerical matching of 1996 and 1997 records by national liaison agency

Solution: Train ML algorithms to learn identity definition

Schematic presentation of overall approach: Phase I



Database Year 1

Database Year 2

Classification: Identifcation of all matches using heuristics and clerical review

Training data

Test data

Record pair comparison

FADN-Variables -> more than 50 features (farm characteristics)

Support vector machines

Random Forests

Neural Networks

Train classification models

Model evaluation

Selected Model M_Y1_Y2

Neural Network

Year 2

Year 3

Year 4

Year N

Database with all direct year-to year matches identified

Results of the matching process: Phase I)

Performance measures

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}$$

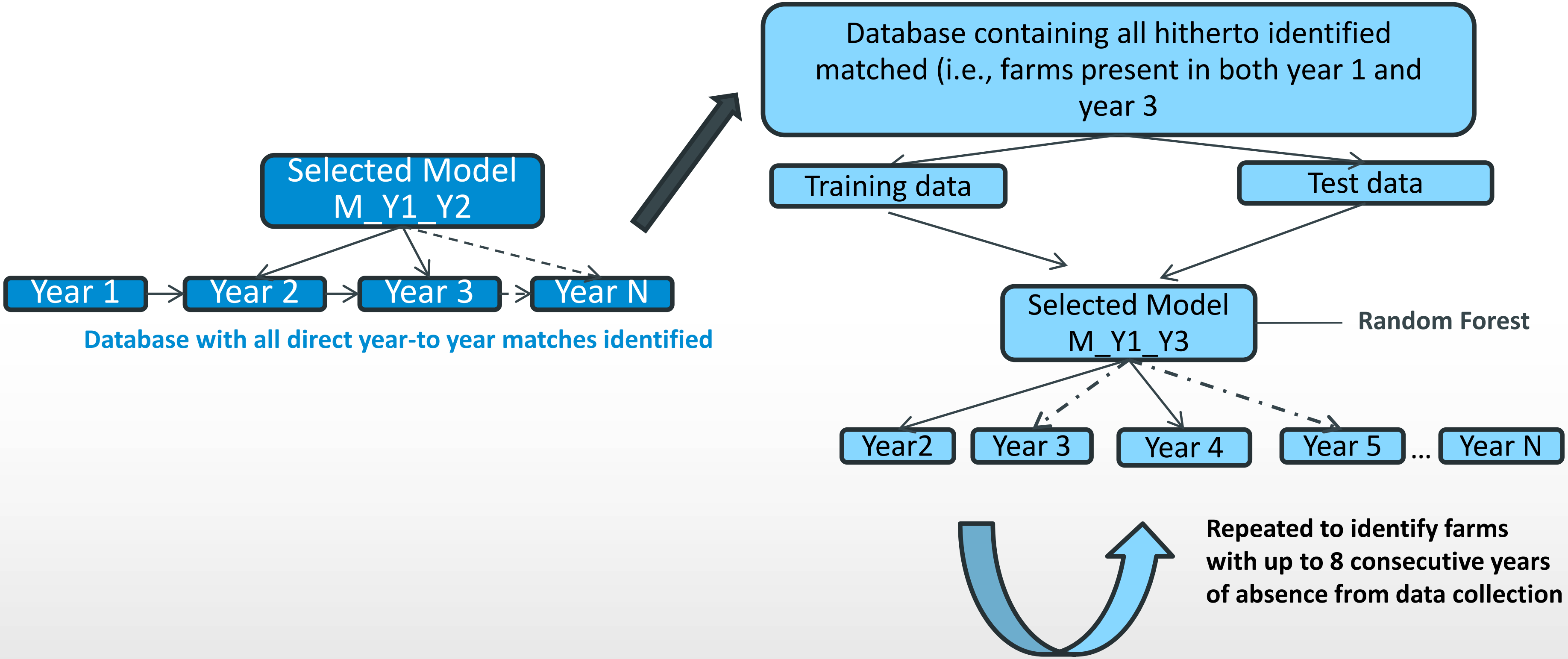
‘true positive’ (TP), ‘true negative’ (TN), ‘false positive’ (FP), ‘false negative’ (FN)

All tested algorithms performed extremely well on year to year matching

Algorithm	tn	fp	fn	tp	F1	Precision	Recall
Linear SVM	9747055	8	9	5366	0,99842	0,99851	0,99833
Random Forest	9747061	2	24	5351	0,99758	0,99963	0,99553
Neural Net	9747059	4	10	5365	0,99870	0,99925	0,99814

Schematic presentation of overall approach

Database with all direct year-to year matches identified



Database containing all hitherto identified matched (i.e., farms present in both year 1 and year 3

Training data

Test data

Selected Model M_Y1_Y3

Random Forest

Year2

Year 3

Year 4

Year 5

…

Year N

Selected Model M_Y1_Y2

Year 1

Year 2

Year 3

Year N

Database with all direct year-to year matches identified

Repeated to identify farms with up to 8 consecutive years of absence from data collection

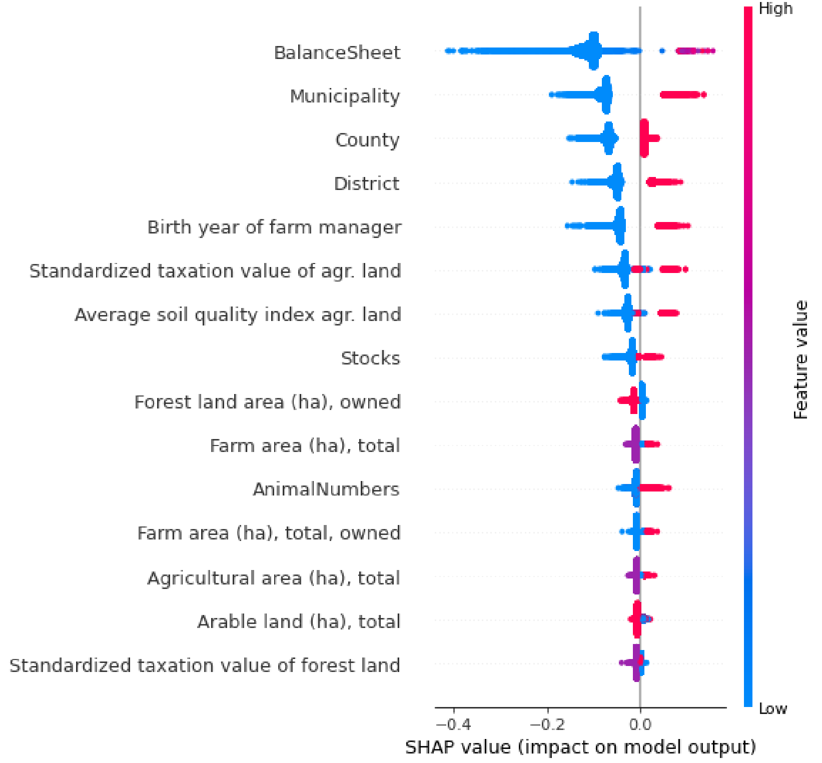
Results of the matching process: Phase II

Random forest models were best suited to identify matches over longer time spans (i.e., if observational data is missing for intermediate years)

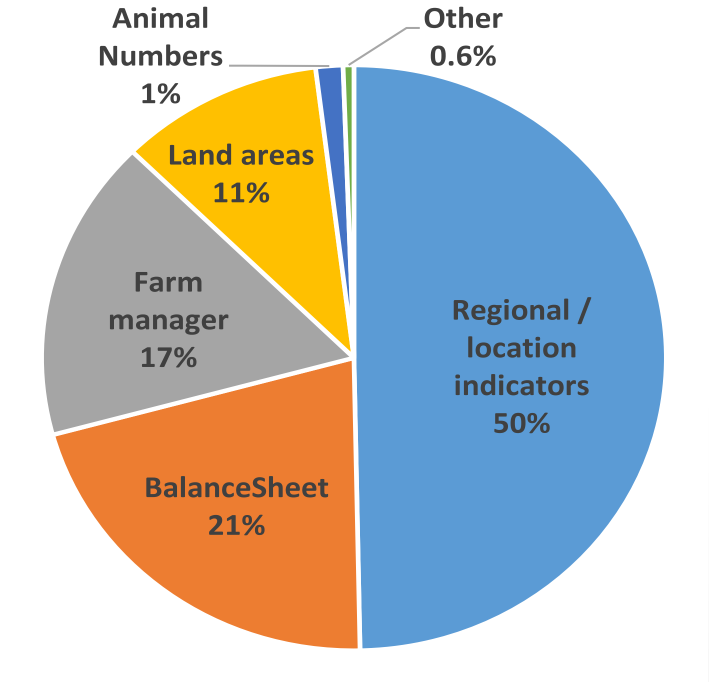
years skipped	F1	Precision	Recall
1	0,97890	0,99957	0,95907
2	0,97295	0,99881	0,94840
3	0,96523	0,99785	0,93467
4	0,95185	0,99698	0,91062
5	0,93714	0,99404	0,88640
6	0,90598	0,98915	0,83570
7	0,90734	0,99089	0,83679
8	0,88815	0,99299	0,80333

- Postprocessing (deduplication of matches) essential

Matching criteria importance



Note: Only the 15 most important features are shown



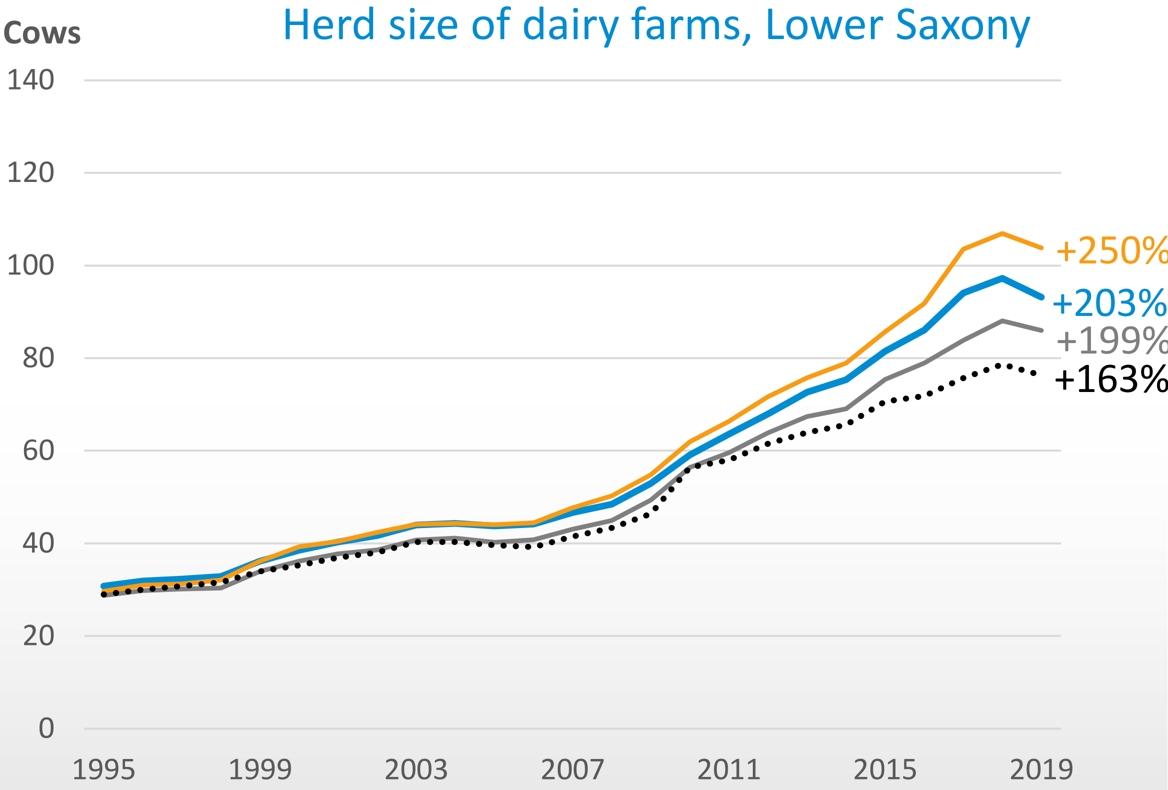
Global feature importance by theme

Impact of different farm identity definitions on sample characteristics (selected results)

	Number of farms present for 25 years
— New identifier	1948
— New identifier, plus farm manager must remain identical	1011
— New identifier, plus legal type must remain identical	1610
······ Original identifier	1004

Impact of different farm identity definitions on sample characteristics (selected results): Herd size

	Number of farms present for 25 years
━ New identifier	1948
━ New identifier, plus farm manager must remain identical	1011
━ New identifier, plus legal type must remain identical	1610
⋯ Original identifier	1004



Conclusions

- Identifier definition can impact sample characteristics
- Identifier available in secondary data may not necessarily match the appropriate definition of 'identity' with respect to the research question

➢ We therefore recommend that studies using longitudinal data provide a more explicit description of how 'identical' is defined for their analysis.

Huge potential of ML algorithms to link farm records across time

➢ and across databases! (c.f. proposed new UUID in EU agr. statistics)

➢ data protection -> privacy preserving linking algorithms!

Some questions to you:

Who is providing the farm identifier (statistical office, liaison agency, ministry)?

How is the identifier defined? Are there guidelines on when an new identifier should be assigned?

How large is the annual turnover (% of new farms in sample)

Do you have farms which are absent from the FADN for one/more years and then return (with the same identifier?)?

Thank you for your attention!

frank.offermann@thuenen.de
Thünen Institute of Farm Economics