MEF4CAP

Case : Federated learning across multiple data stations

Pacioli 28 meeting October 4, 2023 Ptuj, Slovenia

Daoud Urdu, Marcel van Asseldonk, Anand Gavai

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101000662.

Study context: GDPR by-design

- **GDPR** compliance is an essential motivation to apply federeated learning (data train).
- From a compliance perspective: individual farmer data cannot be shared without **consent**
- From a technical perspective: the proposed setup allows data to **interoperate** in a GDPR compliant manner. The raw source data is made **autonomously accessible** for analysis purposes and the user retrieves only result summaries
- The study focuses on the **dairy sector** by enriching data from the FADN with data that **goes beyond the current variables**

Partners who contribute to this demonstration case

- Project partners as FADN data stations
- The data nodes were hosted by partners across three FADN Data stations in
  - the Netherlands (WEcR, Wageningen Economic Research);
  - Poland (IAFE-NRI, Institute of Agricultural and Food Economics);
  - Ireland (TEAGASC, the Agriculture and Food Development Authority).
- Work in progress: figures need to be weighted and further regression analysis

Observation of the past few days

- Interoperability is mentioned a few times
- Machine readability is important
- “What does each variable mean?” Semantics is needed
- Data protection – privacy preserving algorithms are needed

Objective, approach and research questions

Information case-country context: design of the case “Federated learning across multiple data stations”

- Explore technical opportunities to link national datasets for policy evaluation
- Expand the traditional economic analysis of CAP (Common Agricultural Policy) to a broader set of economic, environmental and social objectives

Design oriented approach: privacy preserving federated learning infrastructure

This demonstration case comes to provide an answer to the following questions:

- Is it feasible to set-up a FL approach with FADN Liaison Agencies?
- Is it feasible to unlock distributed data collected by FADN Liaison Agencies which go beyond the current FADN Variables?

What technology developments are analysed?

Towards Federated Learning

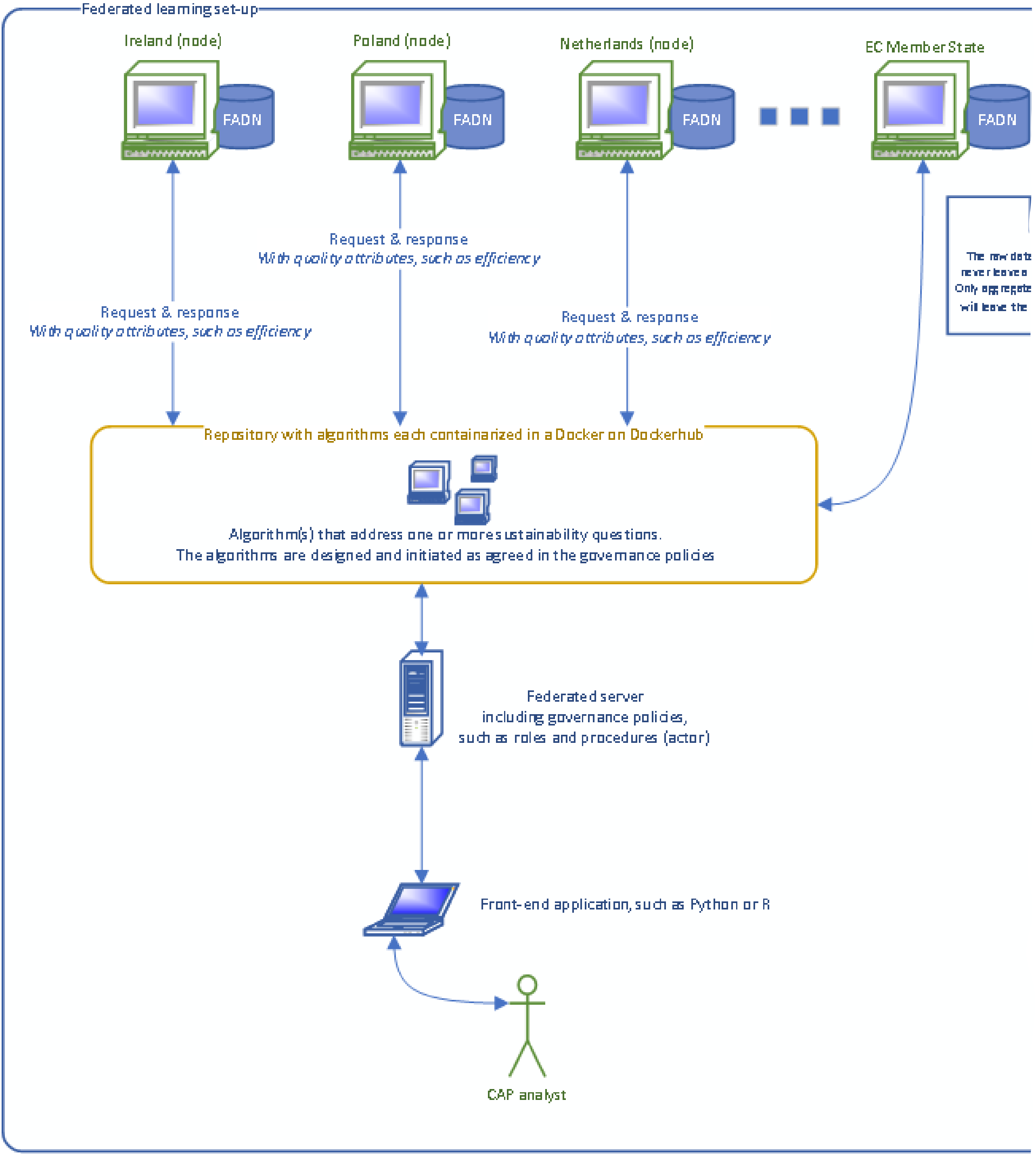
Data Station:
- Provides FAIR access to data and metadata
- Allows train (model to access and interface with data)

Train:
- Interacts with data (these are models that processes data including analysis)

FAIR funder cycle
FAIR data station
Personal Health Train

Data Gateway:
- Authorization and
- Authentication with restricted access to data

Tracking System:
The routing of models and transport infrastructure


Federated learning set-up
Ireland (node)
Poland (node)
Netherlands (node)
EC Member State
FADN
FADN
FADN
FADN
The raw data
never leaves
Only aggregate
will leave the
Request & response
With quality attributes, such as efficiency
Request & response
With quality attributes, such as efficiency
Request & response
With quality attributes, such as efficiency
Repository with algorithms each containarized in a Docker on Dockerhub
Algorithm(s) that address one or more sustainability questions.
The algorithms are designed and initiated as agreed in the governance policies
Federated server
including governance policies,
such as roles and procedures (actor)
Front-end application, such as Python or R
CAP analyst

Conceptual architecture

The federated set-up for this demonstration case

Overall setup

Reusable components

A **central server** that coordinates communication with nodes.

One or more **nodes** that execute algorithms (Docker images)

Organizations (EC Member states) that are interested in collaboration with each other

Collaborations (FADN or FSDN) between organizations

Users, such as policy analysts, that instruct the nodes to execute algorithms

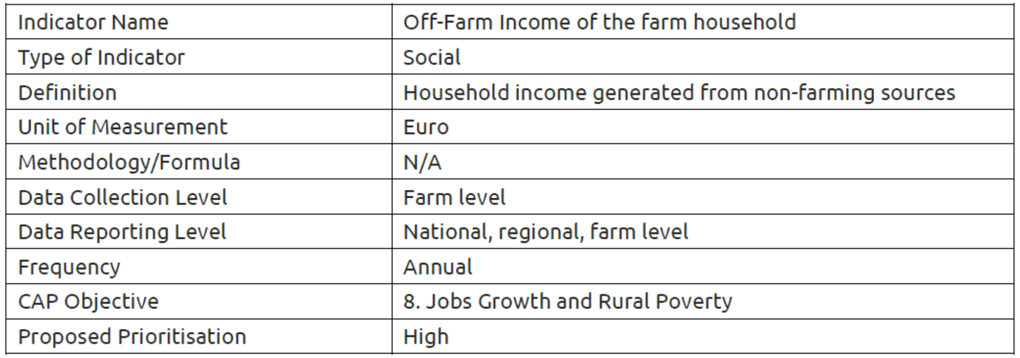
A **Docker registry** that functions as a database of algorithms

Scope

Off-Farm Income for dairy farmers

- Farm structure characteristics
  - Number of dairy cows, total utilised agriculturual area (hectares)
- Farm economic characteristics
  - Farm net income
  - Unpaid labour
  - Total direct payments
- Farmer characteristics
  - Age
  - Sex
  - Agricultural training

Table 36: Farm Household Off-Farm Income



Indicator Name	Off-Farm Income of the farm household
Type of Indicator	Social
Definition	Household income generated from non-farming sources
Unit of Measurement	Euro
Methodology/Formula	N/A
Data Collection Level	Farm level
Data Reporting Level	National, regional, farm level
Frequency	Annual
CAP Objective	8. Jobs Growth and Rural Poverty
Proposed Prioritisation	High

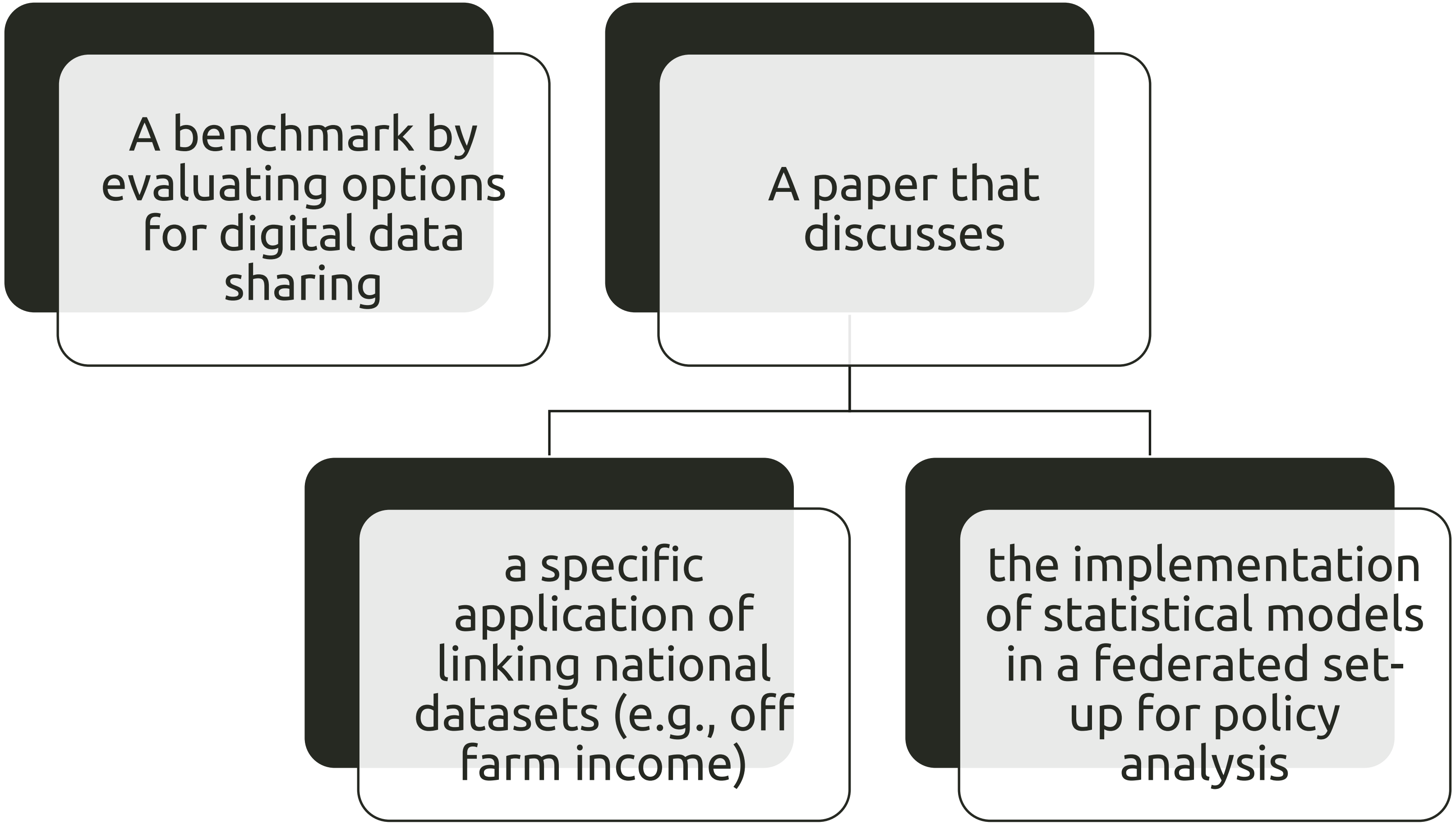
Variables of interest

Table 1 Description of off-farm income and FADN variables of interest

Variable	Unit
Off-farm income	0: no; 1: yes
FADN variables	
Dairy cows	Number of dairy cows
Total utilsed agricultural area	Hectares
Farm net income	Euro per year
Unpaid labour	Annual work units
Total direct payments	Euro per year
Age (years)	Years
Sex	1: male, 2: female
Agricultural training	1: only practical agricultural experience, 2: basic agricultural training, 3: full agricultural training

Expected outcomes

Research paper and project deliverable



A benchmark by evaluating options for digital data sharing

A paper that discusses

a specific application of linking national datasets (e.g., off farm income)

the implementation of statistical models in a federated set-up for policy analysis

Methods and materials

Data Stations, Data Trains and Semantic Rails

Co-design with partners for a federated set-up considering semantic interoperability – Semantic Rails

Design use case and selection of indicators for scoping data on national level – Data Stations

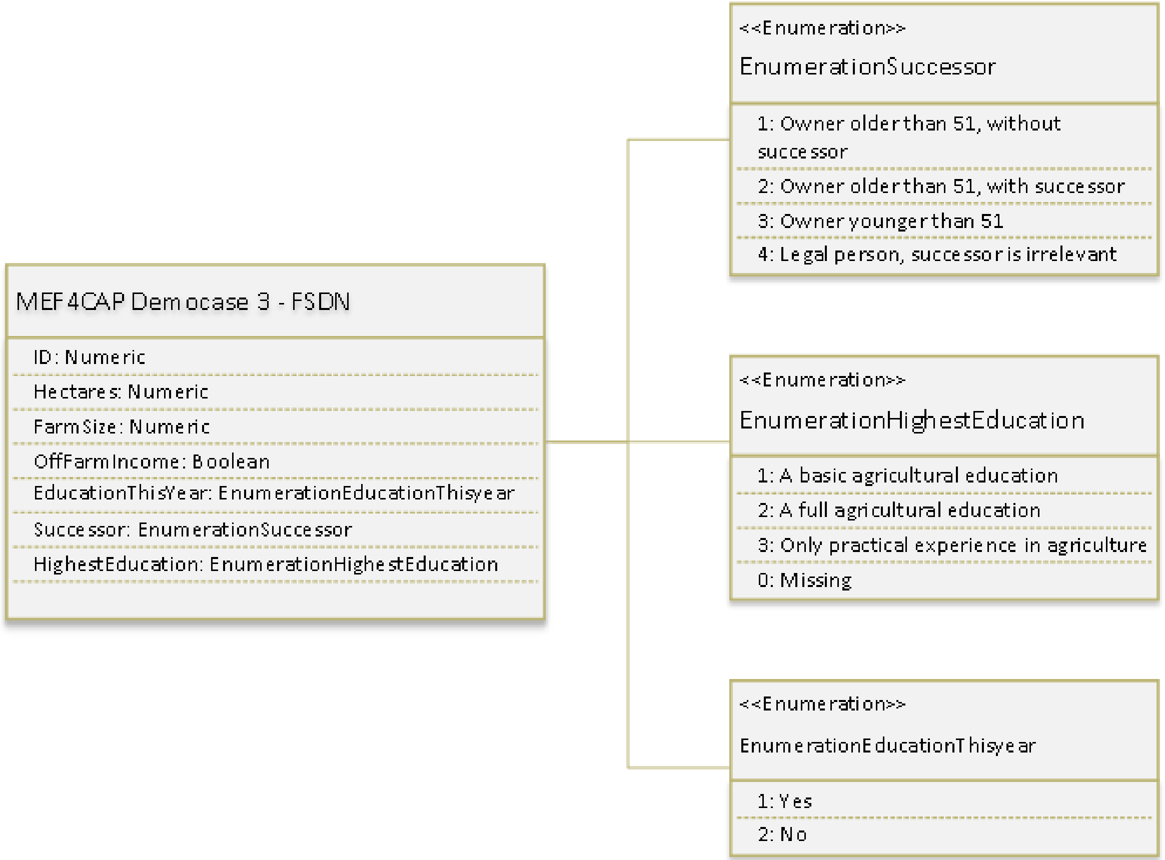
Iterative process of drafting content-relevant questions – Data Trains

FAIR Data Point for interoperability on data station level

To do: Git lab repository for source code on FADN and algorithms for

Initial data model

One table including the variables and several enumerations



MEF4CAP Democase 3 - FSDN

- ID: Numeric
- Hectares: Numeric
- FarmSize: Numeric
- OffFarmIncome: Boolean
- EducationThisYear: EnumerationEducationThisyear
- Successor: EnumerationSuccessor
- HighestEducation: EnumerationHighestEducation

<<Enumeration>>
EnumerationSuccessor

- 1: Owner older than 51, without successor
- 2: Owner older than 51, with successor
- 3: Owner younger than 51
- 4: Legal person, successor is irrelevant

<<Enumeration>>
EnumerationHighestEducation

- 1: A basic agricultural education
- 2: A full agricultural education
- 3: Only practical experience in agriculture
- 0: Missing

<<Enumeration>>
EnumerationEducationThisyear

- 1: Yes
- 2: No

Additional explanation:

- Hectares represents the plot surface and is expressed in hectares.

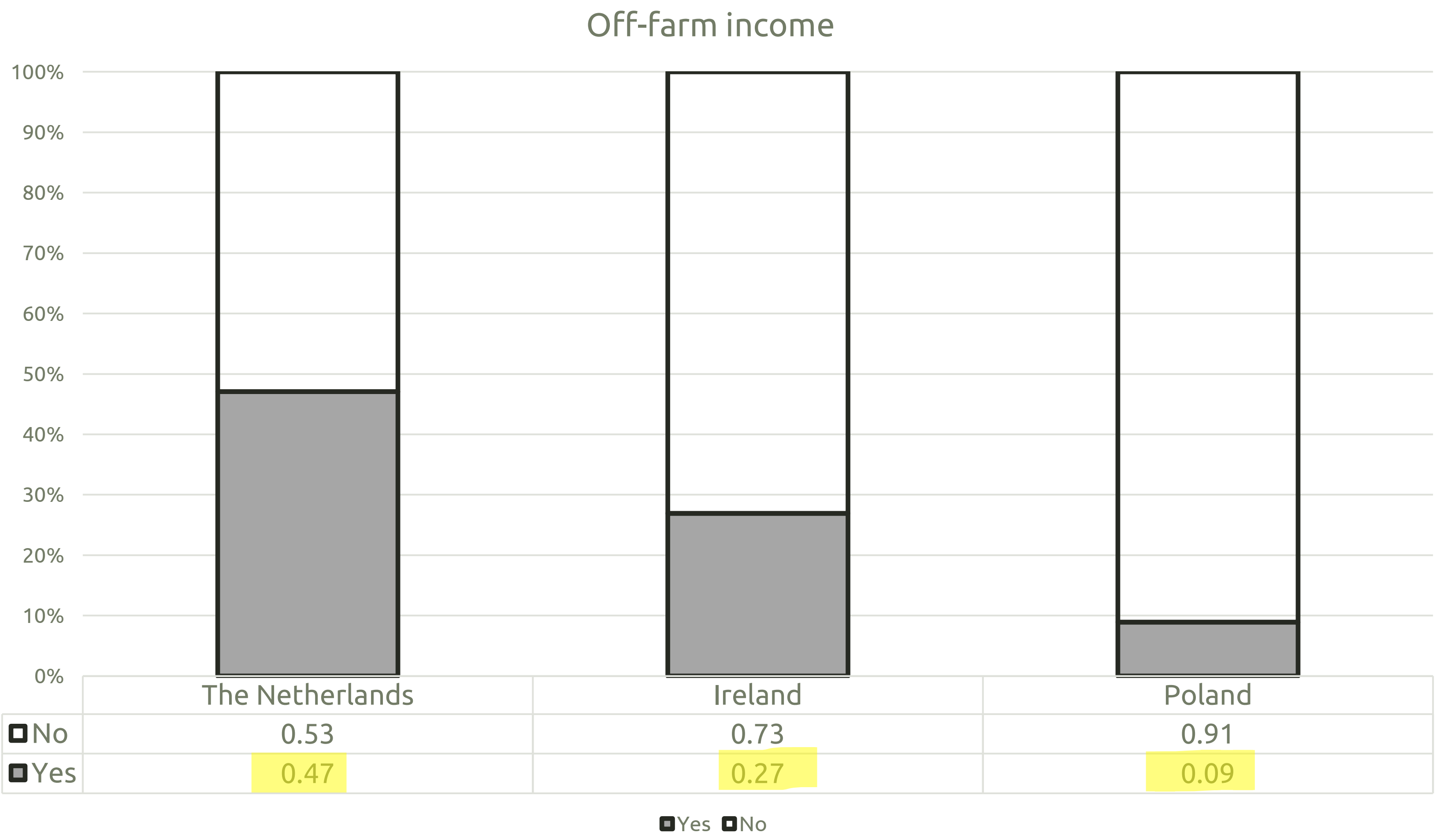
- FarmSize represents the "Standard Yield". This is a standardized measure for the economic size of agricultural companies, based on the yield that is achieved on average on an annual basis per crop or animal produce. The FarmSize is expressed in euros (x1000).

- In case when the value of a variable is unknown then "N/A" is used.

MEF 4CAP

Initial results

Distribution of off-farm income across different data stations



	The Netherlands	Ireland	Poland
□No	0.53	0.73	0.91
▨Yes	0.47	0.27	0.09

Summary statistics

FADN farm structure characteristics of the three data stations

	The Netherlands		Ireland		Poland	
	Yes	No	Yes	No	Yes	No
Off-farm income (%)	47	53	27	73	9	91
FADN variables						
	Mean	Std	Mean	Std	Mean	Std
Dairy cows (number)	110	72	103	64	28	23
Total utilsed agricultural area (hectares)	61	36	75	45	31	24

Summary statistics

FADN farm economic characteristics of the three data stations

	The Netherlands		Ireland		Poland	
	Yes	No	Yes	No	Yes	No
Off-farm income (%)	47	53	27	73	9	91
FADN variables						
	Mean	Std	Mean	Std	Mean	Std
Farm net income (euro)	61,590	100,049	111,152	78,306	40,468	40,781
Unpaid labour (annual work units)	1.49	0.50	1.47	0.61	1.89	0.57
Total direct payments (euro)	25,139	25,628	25,875	16,818	9,506	5,231

Summary statistics

FADN farmer characteristics of the three data stations

	The Netherlands			Ireland			Poland		
	Yes	No		Yes	No		Yes	No	
Off-farm income (%)	47	53		27	73		9	91	
FADN variables									
	Mean	Std		Mean	Std		Mean	Std	
Age (years)	55	9		53	13		45	11	
	Male	Female		Male	Female		Male	Female	
Sex (%)	99%	1%		96%	4%		84%	16%	
	1	2	3	1	2	3	1	2	3
Agricultural training (%)	27%	28%	45%	39%	20%	41%	36%	24%	40%

Initial results: regression analysis

Regression analysis with off-farm income as (binary) dependent variable and FADN characteristics as independent variables of the three data stations

	The Netherlands		Ireland		Poland	
	B	Sig	B	Sig	B	Sig
Dairy cows (number)	0.003	0.165	x	x	x	x
Total utilsed agricultural area (hectares)	-0.003	0.461	x	x	x	x
Farm net income (euro)	0.000	0.795	x	x	x	x
Unpaid labour (annual work units)	-0.166	0.433	x	x	x	x
Total direct payments (euro)	0.000	0.457	x	x	x	x
Age (years)	-0.009	0.410	x	x	x	x
Sex (gender)	0.107	0.916	x	x	x	x
Agricultural training (%)	0.193	0.133	x	x	x	x
Constant	0.116	0.928	x	x	x	x

Concluding results

- The empirical analysis offer a pathway towards achieving semantic interoperability for secure data sharing
- The off-farm income of dairy farmers is closely monitored and assessed
- These findings have wider implications for analysing the effectiveness of measures implemented under the CAP for various business activities and are scalable for other Member States
- This study presents a novel contribution towards enhancing the data-driven decision-making process in the agricultural and food system

Discussion points

- Data democritzation: use of open standards that lead to flexible settings for the application of artificial intelligence, such as Machine Learning models.
- Limited models in a federated setting, only model is the Generalized Linear Model (Cellamare, 2022)
- Efficiency of the tools and languages, R and Python.

Questions for the audience.

What are your thoughts?

- Are there more indicators that you would like to analyse?
  - What could be the response variable, and which are the independent variables? For example, off-farm income is influenced by education and farm_size
- Your thoughts on a common data model and ontology, open for discussion
- …

MEF4CAP

www.mef4cap.eu

@MEF4CAP

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 101000662. The content of this publication exclusively reflects the author's view and the Research Executive Agency and the Commission are not responsible for any use that may be made of the information it contains.