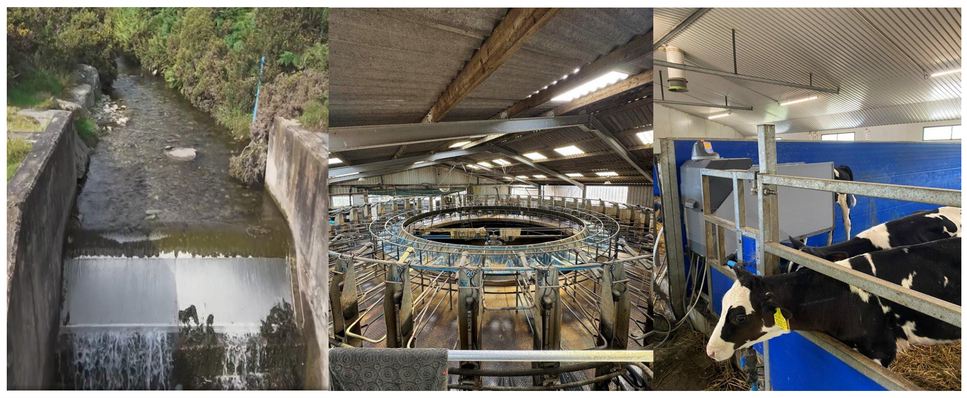
Modelling Greenhouse Gas Emissions from Dairy Farms Through Unsupervised Learning

Jack MacKenzie[1], Vladimir Stankovic[1], Lina Stankovic[1] and Matthias Kuhnert[2]
[1] University of Strathclyde
[2] University of Aberdeen

SUSTAIN
UKRI AI Centre for Doctoral Training in Sustainable Understandable agri-food Systems Transformed by Artificial Intelligence (SUSTAIN)

Motivation

- As global warming is accelerating , the need for green alternatives to energy use is paramount.
- Livestock farmlands occupy approximately 30% of global non-frozen land [1], and the food supply chain is estimated to have produced ~13.7 billion tonnes of CO_2 annually [2].
- Agriculture sector requires **0.5 GtCO2e emission reduction** to achieve UK's net neutrality goals [3].
- Nitrogen fertiliser emissions have decreased over 30 years in the UK, however overall, **Greenhouse Gas (GHG) emissions have risen** due to increased energy intensive farming [4]
- About 80% of land in Scotland is agricultural, thus **raising energy costs are an exponential burden** to the economy as dairy machinery consumes ~80% of all energy in the farm [5]
- Many UK farmers are investing in more sustainable energy sources. However, the means for farmers to assess the impact of their efforts is limited.



- Using machine learning, energy data can be extrapolated automatically, highlighting areas with large amount of GHG emissions, enabling a customised solution to reduce operational costs, GHG emissions and contribute to Britain's carbon net neutrality goals by 2050 [6].

Greenhouse Gas Models

Developed models such as Cool Farm Tool (CFT) [7] are a good way to predict the emissions from farmland. CFT features multiple modules to assess GHG emissions of individual aspects of a farm such as crop fields or livestock on an annual basis. It is possible to expand upon the energy calculations by developing a more detailed analysis on energy consumption through the assessment of seasonal variations and specific equipment usage with Machine Learning.

In the UK, there has been a substantial decrease in CH_4 and N_2O emissions but an increase in CO_2 emissions within the agriculture sector affected by a variety of factors including:

- **Soil type** effectiveness for a chosen crop
- **Crop Rotation** effectiveness for efficient carbon sequestration for the desired crop.
- **Fertiliser usage** and its organic or synthetic origins.
- **Feed Quality** impacting enteric fermentation.
- **Energy imports** from the grid increases emissions indirectly due to non renewable sources; dairy farms have high GHG emissions due to specialised machinery [8].

CFT Model Based Calculations	Data Driven ML Calculations
• Lengthy data collection process	• Utilises sensor data sources such as smart metre data, weather and easily obtainable data including farm size, number of cows
• Single, generalised weighted carbon factors , less suitable for nonlinear relations.	• Multiple weighted carbon factors dependant on daily profiles.
• Useful detail on systems already installed.	• Improved Linearity of energy/features with carbon
• Likely does not recommend alternative energy sources	

Methodology

Framework

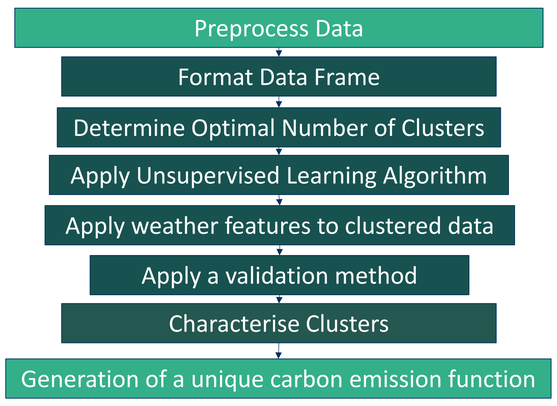


Figure 1, Flow chart of methodology

Data Preprocessing
- All data was pre-processed to remove days with NaN values
- Data was converted into import – export
- Data frames were labelled with weather features and carbon generation

Weather Source – Open Metro API [9]
- **Precipitation mm –** Daily precipitation was acquired and a lag of 0-3 days was implemented.
- **Temperature ºC -** Maximum, minimum and mean temperatures was recorded .
- **Wind Speed km/hr –** Wind speeds from 10 metres.

Carbon Data – Carbon Intensity API [10]
- **Regional Carbon emissions** from the UK grid system in half hour intervals

K-Means Algorithm [11]
Randomly placing centroids around the data based on the number of clusters k, the Euclidian distance between each of the points per centroid is calculated, where the points with the smallest distances become a part of that centroids cluster.

Optimal Number of Clusters
- Essential for estimating the maximum number of clusters before they become over segmented

Elbow Plot [11]
- Provides a prediction of the best number of clusters based off the Within-Cluster Sum of Squares (WCSS, or inertia) [10] which provides a balance of computational power and identifiable features by selecting the elbow (e.g. in Fig.2, $k_{optimal}$ = 2 - 4)

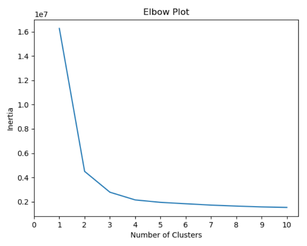


Figure 2, Elbow Plot of consumed energy from Beckside Farm

Statistical Analysis

Key Statistics of clusters

Statistic	K = 3	K = 4	K = 8
David-Bouldin Index (DBI)	0.7584	0.9132	1.484
Calinski-Harabasz Index (CHI)	3451	3116	1797
Effect size of Precipitation with 1 day lag (η^2)	0.2195	0.2412	0.2472

- **Why K = 3?**
  - DBI and CHI conform with the elbow plot and produce distinctly unique clusters
- **Why K = 8?**
  - η^2 demonstrates an increase in weather dependency of clusters, some have similar shapes with varying magnitudes.

Data agnostic cluster quality metrics don't necessary provide optimal k value due to the complexity of external influential factors. This is confirmed visually through the cluster profiles.

Case Study - Beckside Dairy Farm
- Beckside has 3 milking rounds throughout the day with ~800 cows
- Beckside features onsite hydroelectric energy generation
- Import & export energy (kWh) data was gathered between 01/01/21 – 29/11/24

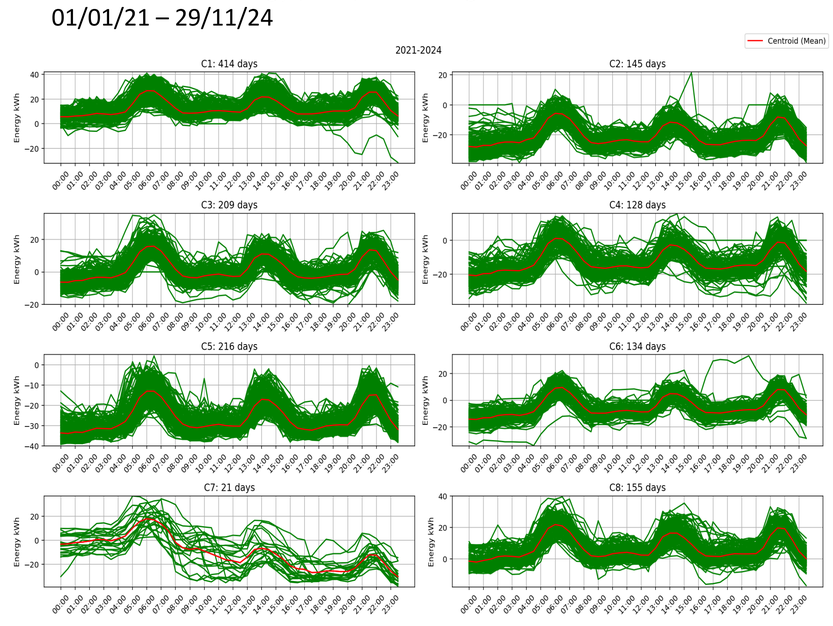
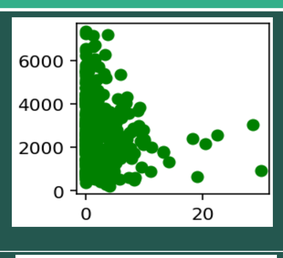
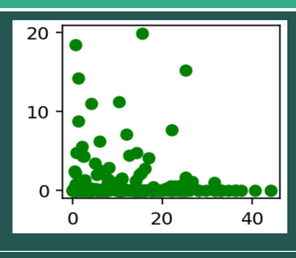
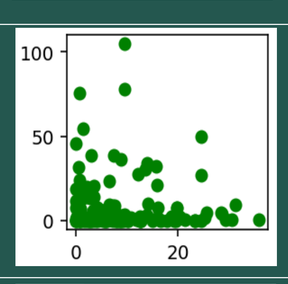
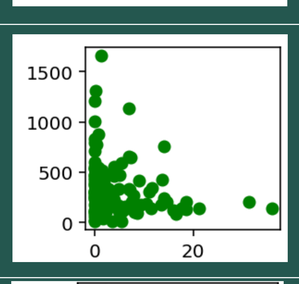
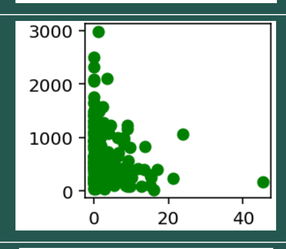
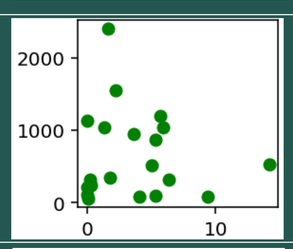
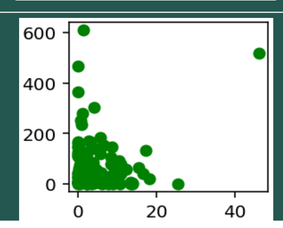
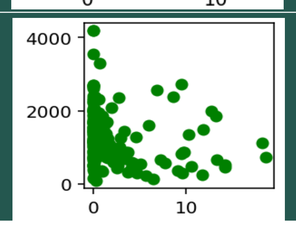


Figure 3, Clusters of energy consumption

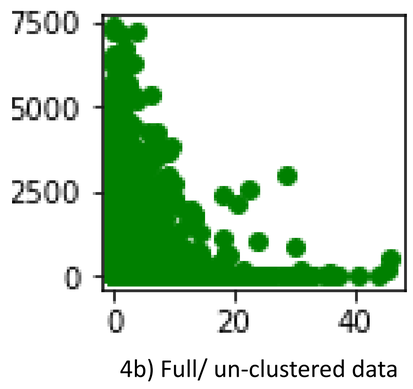
Cluster	Milking round 1 Centroid kWh	Milking round 2 Centroid kWh	Milking round 3 Centroid kWh
1	26.81	19.16	25.64
2	-5.66	-14.04	-8.11
3	15.56	8.49	13.57
4	1.13	-5.84	-1.10
5	-13.03	-20.34	-15.05
6	8.94	2.28	7.96
7	18.00	-7.68	-12.03
8	21.81	14.03	19.61

Towards Quantification of CO_2 Emissions

Figure 4, Carbon kgCO_2 (y) against 1 day lag precipitation mm (x)

Cluster	Carbon kgCO_2 (y) Against 1 Day lag Precipitation mm (x)	Cluster	Carbon kgCO_2 (y) Against 1 Day lag Precipitation mm (x)
1		5	
2		6	
3		7	
4		8	

4a) Clustered Data



4b) Full/ un-clustered data

- There is a clear inverse relationship between carbon emissions and precipitation on farms using onsite – hydro generation.
  - With min mm of rain - 0 mm
  - With max mm of rain – 46.2 mm

Each Farm has unique CO_2 footprint which is lower than current models would calculate
Current GHG models would not capture an emissions factor as granular as a data driven approach.

Future Development

- Carbon emissions to estimate GHG emissions due to energy consumption to establish a footprint.
- Adapt model onto other dairy farms with alternative onsite renewables
- Provide second stage of clustering by grouping together load profiles of multiple farms to capture commonalities

Conclusion

- Clustering produces more linear relations to carbon – emission factor more accurate
- Can create multiple Carbon Emission factors depending on what cluster the day fits into.
- All Cluster milking rounds are unique , showcasing that k = 8 is a suitable number of clusters
- Current GHG models over estimates the CO_2 emissions due to the energy consumption as they do not consider on/off site renewable energy generation

References

[1] M. Herrero et al, (2013) production, feed efficiencies, and greenhouse gas emissions from global livestock systems, Proc. Natl. Acad. Sci. U.S.A. 110 (52) 20888-20893,
[2] Shabir, I. et al. (2023) 'Carbon footprints evaluation for sustainable food processing system development: A comprehensive review', *Future Foods*, 7, p. 100215.+
[3] IEA. Global energy-related CO2 emissions by sector in 2020 and 2050 (2020).
[4] Department for Environment Food & Rural Affairs. Agri-climate report 2022.
[5] Vavouris, A., Stankovic, L. and Stankovic, V. (2025) 'A non-intrusive load monitoring-enabled framework for load scheduling in the dairy industry',
[6] House of Commons Library (2025) The UK's plans and progress to reach net zero by 2050. Research Briefing No. 9888, 6 November. London: House of Commons.
[7] Cool Farm Alliance. (2025) 'Cool Farm Tool and Platform - Technical Method Description - CFP version 3 and CFT version 2'.
[8] Thoma, G. *et al.* (2013) 'Greenhouse gas emissions from milk production and consumption in the United States: A cradle-to-grave life cycle assessment circa 2008', *International Dairy Journal*, 31, pp. S3–S14
[9] P. Zippenfenig, "Open-meteo.com weather api," 2023.
[10] National Energy System Operator (2026) Carbon Intensity API.
[11] R. S. Mollel et al, "Characterising Namibian Residential Load Profile Archetypes Using Self-Organising Map (SOM)-Based Temporal Segmentation and Seasonal Clustering," *2025 International Conference on ETNCC*, Windhoek, Namibia, 2025, pp. 1-9.

This work was supported by UKRI [grant ref: EP/Y03063X/1]