**Steps for the Simulations**

1. Synthetic Data by running the MILP model, Electroscape.
   1. Time step 1 hour Model assumption, April 1, 2019, to March 31, 2020.
   2. Number of households 400.

**Inputs:**

* 1. Underlying demand profile Time series kW
  2. Underlying solar profile Time series kW/kWp
  3. Tariff schemes: A:Fi-Fe, B:Ti-Te, C:Ti-Fe and D:Fi-Te (From 1 July 2022– 30 June 2023)

PV\_Only

|  |  |
| --- | --- |
| pv\_min\_capacity | 1 |
| pv\_max\_capacity | 20 |
| pv\_fixed\_cost | 3000 |
| pv\_cost\_per\_kW | 630 |
| bess\_min\_capacity |  |
| bess\_max\_capacity |  |
| bess\_fixed\_cost |  |
| bess\_cost\_per\_kWh |  |

PV\_BESS

|  |  |
| --- | --- |
| pv\_min\_capacity | 1 |
| pv\_max\_capacity | 20 |
| pv\_fixed\_cost | 3000 |
| pv\_cost\_per\_kW | 630 |
| bess\_min\_capacity | 3 |
| bess\_max\_capacity | 20 |
| bess\_fixed\_cost | 0 |
| bess\_cost\_per\_kWh | 1300 |

**Output:**

* PV\_size only
* PV+BESS size

1. Scenarios
   1. A,B, C used for training and validation
   2. D (FiTe): New unseen, serves as a generalisation test
2. The following items provide a framework for the surrogate modelling process presented in the study:
   1. Using the entire 8760 hourly time to train, validate and test the neural networks.
   2. Data Preprocessing:
      * + 1. Scaling and normalising.
          2. Handling missing data.
          3. Identifying and treating outliers.
   3. The training dataset can be further split into sub-datasets for training and validation according to a specific percentage.

* Training: 300 households (75% of the dataset).
* Validation: Use 100 households (25% of the dataset).
  1. Additionally, exploring steps such as sequential and clustering methods. Then, based on these insights, the preprocessing and data splitting steps (b and c) are repeated.

1. Neural network-based surrogate model training for scenarios A, B and C:
   * 1. PV-Only, B) PV+BESS.
   1. Architecture of NNs:
   2. **Multi-Layer Perceptron (MLP)**
   3. **Long Short-Term Memory (LSTM)**
   4. **CNN**
   5. Determination of hyperparameters
      * 1. Hyperparameters Optimisation: Finding a set of model hyperparameters, such as number of epochs, hidden layers, activation functions, learning rate, etc., **to enable the model to better capture the relationship between the inputs and outputs.**
2. Testing on Scenario D
   1. Scenario D (FiTe) -unseen dataset. Used to test the trained neural network (Generalisation capability of the trained model)
3. Evaluating models with **error metrics** like MAE, RMSE, and **R²** for both:
   1. Validation (A–C)
   2. Test (D)

Next Step in the following days:

1. Generate the data set for D (FiTe): both PV-Only and PV+BEES.
2. Training, validating and testing the neural network. Using all the data from the time series without feature engineering.

Note: A further approach is to use the sample dataset for training and build an additional dataset for testing.

**Next Requirements**

1. **PV+BESS Data for Lisa’s Model**
   * Required: Output data for PV+BESS configuration - A:Fi-Fe, B:Ti-Te, C:Ti-Fe (BESS cost per kWh:1300)
   * Available: BESS cost per kWh at **300, 400, and 500 AUD-** A:Fi-Fe, B:Ti-Te, C:Ti-Fe. It is not available D:Fi-Te.
2. **Run the same set of scenarios (A, B, C)**
   * But using **BESS cost per kWh = 1300 AUD** (standard benchmark)
   * Ensure consistency across all Fi-Fe, Ti-Te, and Ti-Fe tariff structures
3. I have scenarios A:Fi-Fe, B:Ti-Te, C:Ti-Fe. (250 households) tariff 2024-2025)

It is not available D:Fi-Te.

1. Working in different neural network configurations, MLP-LSTM and CNN.
   * + 1. Use each scenario (A:Fi-Fe, B:Ti-Te, C:Ti-Fe), changing the tariff 0%, 5%-+5%. Having nine scenarios, 3 for each neural network individually,
       2. Using all the scenarios together for the neural networks.
       3. Combined all the neural networks.

**TIMELINE**

A chart with text on it

AI-generated content may be incorrect.

**Revise Methodology:**

**Step 1**

Data acquisition

**Step 2**

* pre-processes raw features based on domain knowledge to form an improved candidate feature set.
* Identifies statistically significant features from the candidate feature set, using filter method. Removing irrelevant features and removing redundant features. (More common Chi-squared test,, correlation,entropy-based criteria)
* Decides on the final feature set used for the model

Feature sequential, Clustering algorithms to segment consumers and obtain the representative load patterns based on diurnal load profiles. (Discrete wavelet transform)

* The proposed method uses discrete wavelet transform (DWT) to extract features from daily electricity consumption data.
* The extracted features are reconstructed using a statistical method, combined with Kendal correlation coefficient.
* Important features: Create a vector with the most important features (Ex. P-values)
* symplectic geometry model decomposition (SGMD) to decompose the feature data and obtain multiple feature sub-sequence?

**Issue:**

* A variable with low correlation to the model output can significantly enhance performance when combined with other features.
* two variables that have high correlations (or anti-correlations) do not necessarily mean that they would not complement each other when being used in a model. Ref. Liang Zhang, Jin Wen, 2019.

**Step 3**

Feature sequences dimensionality reduction based on the KPCA

Why: KPCA is a nonlinear data processing method based on the kernel function, which is one of the improved PCA methods, capturing non-linear structure of the data. (Simin Peng, 2024)

Issue: To select Kernel function.

* PCA: The feature after dimensionality reduction cannot effectively represent the structure of the original data.

**Step 4**

Model training. The training dataset is used for the model training, and the optimal hyper-parameters.

* I need to check the constraints from MILP
* MILP need to be preserved in the NNs-based representation

How:

* Add penalty terms to your loss function based on constraint violations.

**Step 5**

Error analysis and model verification. Different evaluation indices are used to judge the prediction results. Compared with other prediction methods.

**Phase 1: Data Preparation**

1. Clean and align all hourly data (load, solar, tariffs)
2. Generate aggregated and sequence-based input features
3. Compute binary masks for PV/BESS deployment
4. Compute SNR and installation rates

**📌 Phase 2: Surrogate Modeling**

1. Train base models: ANN, CNN, LSTM with aggregated and sequence inputs
2. Evaluate model performance (MAE, R², visual inspection)
3. Compare individual models and ensemble predictions

**SNR:**

Noise definition (for PV and NPV)

* 1. ~~NaN/Nulls (MILP does not present NaN/Nulls)~~
  2. zero values (MILP presents some zero values values)
  3. Values that differ from the MILP analysis more than a given percentage (5% error proposed)