A dataset for residential buildings energy consumption with statistical and machine learning analysis

We present a new dataset for energy consumption of residential buildings and examines the impact of residential buildings' eight input variables (Building Size, Floor Height, Glazing Area, Wall Area, window to wall ratio (WWR), Win Glazing U-value, Roof U-value, and External Wall U-value) on the heating load (HL) and cooling load (CL) output variables.

Methods and statistical analysis results using SPSS statistical software:

Data Exploration

The simulated buildings were generated using the IES<VE> simulation software (Available: https://www.iesve.com/)

The probability distributions of all the input and output variables using Histogram.
Using the file (Dataset.sav) and do the following steps:

1. Go to Analyze menu
2. Choose Descriptive Statistics – then Descriptive
3. Go to charts
4. Click on Histogram
5. Then Ok

The scatter plots for each of the input variables with each of the two output variables
Using the file (Dataset.sav) and do the following steps:

1. Go to Graphs menu
2. Choose Chart builder – then Scatter
3. Drag on the Y-axis CL
4. Drag on the X-axis one of the input variables
5. Repeat the steps for all the 8 inputs with CL
6. Repeat the steps for all the 8 inputs with HL in the Y-axis
7. Then Ok

Statistical Analysis using SPSS statistical software

The Spearman rank correlation coefficient
Using the file (Dataset.sav) and do the following steps:

1. Go to Analyze menu
2. Choose Correlate – then Bivariate
3. Go to variables
4. Click on Spearman
5. Then Ok

Machine Learning-Based Analysis using SPSS statistical software:

Multiple Linear Regression (MLR) Analysis

The normal P-P plot of the standardized residual for dependent variables CL and HL

Using the file (Dataset.sav) and do the following steps:

1. Go to Analyze menu
2. Choose Regression – then Linear
3. Go to dependent variables and choose CL
4. Go to independent variables and choose all the eight input variables
5. Go to Plots – click on Normal P-P plot
6. Choose the ZRESID to be on Y axis and ZPRED on the X-axis
7. Click on fit line at total
8. Then Ok
9. **Repeat all the step with dependent variables and choose HL**

The scatter plot of the regression standardized residual for CL and HL

Using the file (Dataset.sav) and do the following steps:

1. Go to Analyze menu
2. Choose Regression – then Linear
3. Go to dependent variables and choose CL
4. Go to independent variables and choose all the eight input variables
5. Go to Plots – click on Scatter regression standard
6. Choose the ZRESID to be on Y axis and ZPRED on the X-axis
7. Click on fit line at total
8. Then Ok
9. **Repeat all the step with dependent variables and choose HL**

The 10-fold Cross validation (CV)

Using the file (Dataset -CV-10-1-CL.sav) and do the following steps:

1. Go to Data menu - Choose Select cases - Click on Random sample of cases (Approximately 90%)
2. We will find a new column called (Filter) change it Sample
3. Go to again to Data menu - Choose Select cases
4. Click on if conditional Sample=1 (to be sure the all the selected are 90%)
5. Go to Analyze menu - Choose Regression – then Linear
6. Go to dependent variables and choose CL
7. Go to independent variables and choose all the eight input variables
8. Choose from methods: Stepwise – then OK
9. Go to Data menu – choose All cases
10. Go to Transform – Compute value – Target=Predict and write the equation according to the output

11. Then Ok – a new column generated called (Predict) **copy the first 10%** from these values to your Excel sheet to calculate the required performance measures
12. Repeat the previous steps using the file (Dataset -CV-10-2-CL.sav) and after press Ok – a new column generated called (Predict) **copy the second 10%** from these values to your Excel sheet to calculate the required performance measures.
13. Repeat the previous steps using the file (Dataset -CV-10-3-CL.sav) and after press Ok – a new column generated called (Predict) **copy the third 10%** from these values to your Excel sheet to calculate the required performance measures.
14. Repeat the previous steps using the file (Dataset -CV-10-4-CL.sav) and after press Ok – a new column generated called (Predict) **copy the fourth 10%** from these values to your Excel sheet to calculate the required performance measures.
15. Repeat the previous steps using the file (Dataset -CV-10-4-CL.sav) and after press Ok – a new column generated called (Predict) **copy the fifth 10%** from these values to your Excel sheet to calculate the required performance measures.
16. Repeat the previous steps using the file (Dataset -CV-10-6-CL.sav) and after press Ok – a new column generated called (Predict) **copy the sixth 10%** from these values to your Excel sheet to calculate the required performance measures.
17. Repeat the previous steps using the file (Dataset -CV-10-7-CL.sav) and after press Ok – a new column generated called (Predict) **copy the seventh 10%** from these values to your Excel sheet to calculate the required performance measures.
18. Repeat the previous steps using the file (Dataset -CV-10-8-CL.sav) and after press Ok – a new column generated called (Predict) **copy the eighth 10%** from these values to your Excel sheet to calculate the required performance measures.
19. Repeat the previous steps using the file (Dataset -CV-10-9-CL.sav) and after press Ok – a new column generated called (Predict) **copy the ninth 10%** from these values to your Excel sheet to calculate the required performance measures.
20. Repeat the previous steps using the file (Dataset -CV-10-10-CL.sav) and after press Ok – a new column generated called (Predict) **copy the last 10%** from these values to your Excel sheet to calculate the required performance measures.

Repeat all the step with dependent variables and choose HL using the following files:
Dataset -CV-10-1-HL.sav
Dataset -CV-10-2-HL.sav
Dataset -CV-10-3-HL.sav
Dataset -CV-10-4-HL.sav
Dataset -CV-10-5-HL.sav
Dataset -CV-10-6-HL.sav
Dataset -CV-10-7-HL.sav
Dataset -CV-10-8-HL.sav
Dataset -CV-10-9-HL.sav
Dataset -CV-10-10-HL.sav

The mean value of each MLR coefficient over the 10-fold CV iterations are obtained and used for predicting CL and HL (and the equations)

Using the file (Dataset – Cross.sav) and do the following steps:

1. Go to Analyze menu
2. Choose Regression – then Linear
3. Go to dependent variables and choose CL
4. Go to independent variables and choose all the eight input variables
5. Choose from methods: Enter
6. Go to statistics and click on Estimates
7. Then Ok
8. **Repeat all the step with dependent variables and choose HL**

Multilayer Perceptron (MLP) Analysis:

The following three different distributions for the dataset are applied:

(i) 70% to train the NN and 30% to test the NN;

Using the file (Dataset - MLP-CL-70-80-90.sav) and do the following steps:

1. Go to Data menu - Choose Select cases - Click on Random sample of cases (Approximately 70%)
2. We will find a new column called (Filter) change it Sample
3. Go to again to Data menu - Choose Select cases
4. Click on if conditional Sample=1 (to be sure the all the selected are 70%)
5. Go to Analyze menu - Choose Regression – then Linear
6. Go to dependent variables and choose CL
7. Go to independent variables and choose all the eight input variables
8. Choose from methods: Stepwise – then OK
9. Go to Data menu – choose All cases
10. Go to Transform – Compute value – Target=Predict and write the equation according to the output
11. Then Ok – a new column generated called (Predict) copy these values to your Excel sheet to calculate the required performance measures
12. **Repeat all the step with dependent variables and choose HL and use the file (Dataset - MLP-HL-70-80-90.sav)**

(ii) 80% to train the NN and 20% to test the NN;

Using the file (Dataset - MLP-HL-70-80-90.sav) and do the following steps:

1. Go to Data menu - Choose Select cases - Click on Random sample of cases (Approximately 80%)
2. We will find a new column called (Filter) change it Sample
3. Go to again to Data menu - Choose Select cases
4. Click on if conditional Sample=1 (to be sure the all the selected are 80%)
5. Go to Analyze menu - Choose Regression – then Linear
6. Go to dependent variables and choose CL
7. Go to independent variables and choose all the eight input variables
8. Choose from methods: Stepwise – then OK

9. Go to Data menu – choose All cases
10. Go to Transform – Compute value – Target=Predict and write the equation according to the output
11. Then Ok – a new column generated called (Predict) copy these values to your Excel sheet to calculate the required performance measures
12. **Repeat all the step with dependent variables and choose HL and use the file (Dataset - MLP-HL-70-80-90.sav)**

(iii) 90% to train the NN and 10% to test the NN.

Using the file (Dataset - MLP-HL-70-80-90.sav) and do the following steps:

1. Go to Data menu - Choose Select cases - Click on Random sample of cases (Approximately 90%)
2. We will find a new column called (Filter) change it Sample
3. Go to again to Data menu - Choose Select cases
4. Click on if conditional Sample=1 (to be sure the all the selected are 90%)
5. Go to Analyze menu - Choose Regression – then Linear
6. Go to dependent variables and choose CL
7. Go to independent variables and choose all the eight input variables
8. Choose from methods: Stepwise – then OK
9. Go to Data menu – choose All cases
10. Go to Transform – Compute value – Target=Predict and write the equation according to the output
11. Then Ok – a new column generated called (Predict) copy these values to your Excel sheet to calculate the required performance measures
12. **Repeat all the step with dependent variables and choose HL and use the file (Dataset - MLP-HL-70-80-90.sav)**

The obtained NNs to predict CL and HL from the set of 8 input variables and the importance score of the input variables:

Using the file (MLP-CL – new.sav) and do the following steps:

1. Go to Analyze menu
2. Choose Neural Network – then Multilayer Perceptron
3. Go to dependent variables and choose CL
4. Go to Covariates variables and choose all the eight input variables
5. Make the partitions (70 and 30)
6. Go to output tab – click on independent variables importance
7. Go to the Save tab – click on Save predicted
8. Then Ok
9. Repeat with partitions (80 and 20) and (90 and 10)
10. **Repeat all the step with dependent variables and choose HL using the file (MLP-HL – new.sav)**