Determination of Regression Parameters for the Thermal and Energy Components of Smart Homes

Oleh Sinkevych
Department of Radioelectronic and Computer Systems
Ivan Franko National University of Lviv
Lviv, Ukraine
oleh.sinkevych@lnu.edu.ua

Liubomyr Monastyrskyi
Department of Radioelectronic and Computer Systems
Ivan Franko National University of Lviv
Lviv, Ukraine
liubomyr.monastyrskyy@lnu.edu.ua

Bohdan Sokolovskyi
Department of Radioelectronic and Computer Systems
Ivan Franko National University of Lviv
Lviv, Ukraine
bohdan.sokolovskyy@lnu.edu.ua

Abstract—**Determination of the regression parameters of different data sets which consist of the thermal and energy components of smart homes is carried out. Analyzed data were automatically gathered by installed sensors under UK REFIT Smart Home project and published for an open access. In order to carry through the statistical analysis of the given data, the research stages have been divided into three steps. Firstly, the given data have been stored into a specially developed sqlite database. Secondly, the data prepossessing (cleaning, smoothing and resampling) and visualizing have been provided using Python libraries (pandas, seaborn, matplotlib). Then, the cleaned and smoothed data have been used for the statistical regression analysis within which different hypotheses of relations have been investigated. Conducted study of the thermal and energy components of the data has shown a suitability of such data for a forecasting and formulation of the inverse problem of determination of the smart home thermophysical parameters.**

Index Terms—**smart home, home automation, data acquisition, data mining.**

I. Introduction

Nowadays, an evolution of effective intelligent solutions for smart grid technologies become more and more important because of raising problems of smart home energy savings [1], providing comfortable dwelling conditions [2] and rapid development of embedded sensor systems [3]. Variety of suggested methods and solutions have being increased during the last decade: implementation of the artificial neural networks (e.g., recurrent neural networks using well-designed frameworks like tensorflow, caffe, torch, etc) and machine learning algorithms for the embedded system control [4], integration with batteries and renewable energy sources [5], motion and resident's behavior analysis (based on machine learning techniques) for the improvements of system's functioning. Also, solving the energy demand problem during different tariff periods is of a considerable interest among the mentioned challenges [6].

Last but not least problem is an energy disaggregation which implies an application of advanced machine learning algorithms to split aggregated data into single-appliance portions of consumption [7].

In this work, we carry out statistical estimation of regression relations for smart home energy and temperature data on the available open access REFIT Smart Home.

II. Data Description and Preprocessing Techniques

A. Description of the Raw Data

Raw REFIT Smart Home data consist of three main files with structure description, information about dwellings (20 buildings) with installed sensors (measurements of outdoor temperatures, set of room air and corresponding radiator temperatures, gas usage) and sensor readings which were selected for annual period (March, 2014 – March, 2015). In order to simplify the work with big size/high frequency data, special sqlite database has been developed with the use of Python 3.5 and sqlalchemy toolkit.

Gas usage measured in m^3 has been converted to kWh via following formula [8]

$$g_i(kWh) = \left(g_i\left(m^3\right) \cdot c \cdot cv\right) / c_f \,, \qquad (1)$$

where i is the time index, $c = 1.02264$ is the industry standard conversion factor, $cv = 39.3$ is the calorific value and $c_f = 3.6$ is the conversion factor.

Though the gas has been used no only for the compulsory heating process, in this work we have considered the gas to be used for the dwelling heating completely. To remove the ambiguity of which appliance (shower, dishwasher, radiator, etc.) uses certain gas portion the disaggregation problem should be investigated and solved.

The gas usage, external and indoor temperature readings (for each corresponding dwelling space) and surface radiator temperature readings (in accordance with the space) were gathered every 15 minutes and were taken as input data for our research.

B. Data Preprocessing

Necessary preprocessing steps (creation of the database, data cleaning, etc.) have been done to prepare data for a subsequent analysis within which the statistical methods have been used to select the best model.

During the visualization of raw data sets we have found out gaps of missing values for different time periods and large number of 'spikes' which could be caused by anomalies or reading errors. To deal with the gaps we have chosen the state-of-art methods like interpolation techniques (linear, polynomial, spline) among which the simple linear interpolation has been applied

$$x_i = (1-t_i)x_a + t_i x_b,\ t_i = t_0 + i/(n-1), \tag{2}$$

where $x_i \in [x_a, x_b]$ is the gap between x_a and x_b endpoints, n is the number of missed values in a gap.

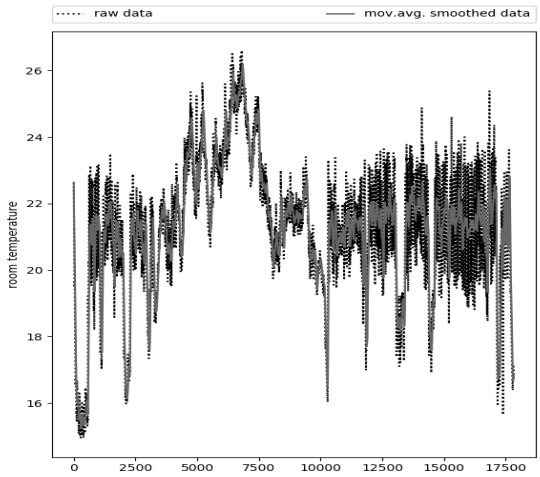

Fig.1. Raw and exponentially smoothed data

The interpolated data were smoothed with using the exponential smoothing algorithm (see Fig. 1)

$$e_1^s = x_i,\ e_{i+1}^s = \alpha x_{i+1} + (1-\alpha)e_i^s, \tag{3}$$

where $\alpha = 0.1$ is the smoothing parameter which has been chosen heuristically.

In order to select proper smoothing method, advanced investigation of the data type, frequency and seasonal components should be taken into account. To simplify this stage we used a simple exponential smoothing method. It has allowed us to reduce small amount of 'spikes'. It should be noted that these 'spikes' can indicate the significant data features which also have to be carefully investigated.

The last step of data preprocessing consists of the data resampling. We have used daily averaging approach when all data samples have been averaged over each day of the selected annual period.

III. Determination of Regression Parameters

A. Calculation of regressions

To study the magnitude of the relations between the gas usage, indoor temperatures, outdoor temperatures and surface radiator temperatures we have built multiple regression models of different degrees in order to find the best fitness. For this stage we have used the following formula

$$\tilde{y}_i = \sum_{j=0}^{n} \beta_{ij} x_i^j + \varepsilon_i,\ x_0^j = 1, \tag{3}$$

where (x_i, y_i), $i = 1,...k$, is the single data observation, ε_i are the model errors, β_{ij} are the regression coefficients, $\tilde{y}_i$ is the predicted value. To estimate these errors the least squares method can be effectively implemented to minimize residual sum of squares (RSS)

$$RSS(\beta_{ij}) = \sum_{i=0}^{k}\left(y_i - \sum_{j=1}^{n} x_i^j \beta_{ij} \right) \to \min_{\beta}. \tag{4}$$

To calculate values of the parameters $\boldsymbol{\beta} = \beta_{i0}, \beta_{i1}, ..., \beta_{in}$ we have used matrix notation

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{5}$$

$$\mathbf{X} = \begin{bmatrix} 1 & \dots & x_1^n \\ \dots & \dots & \dots \\ 1 & \dots & x_k^n \end{bmatrix},\ \mathbf{y} = \begin{bmatrix} y_1 \\ \dots \\ y_k \end{bmatrix}, \tag{6}$$

where $\mathbf{X}$ and $\mathbf{y}$ are the matrix and the vector of observations, respectively.

An estimation of the models fitness has been done with the use of the following formulas (calculations of mean squared error and r^2 value)

$$MSE(\mathbf{y},\tilde{\mathbf{y}}) = \sum_{i=1}^{k} E(\tilde{\mathbf{y}} - \mathbf{y})^2 + \sum_{i=1}^{k} \mathrm{var}(\tilde{\mathbf{y}}), \quad (7)$$

$$r^2(\mathbf{y},\tilde{\mathbf{y}}) = 1 - \sum_{i=1}^{k} (\tilde{\mathbf{y}} - \mathbf{y})^2 \Big/ (\mathbf{y} - mean(\mathbf{y})), \quad (8)$$

where $E(\cdot)$ is the mathematical expectation, $\mathrm{var}(\cdot)$ is the statistical estimation of the variance.

B. Results

Here studying the regression parameters, i.e., relation magnitudes, consists of the selection of proper model degree via a pipeline (consecutive polynomial model fitting with interaction term). Then, based on equations (7), (8) we have estimated the best model type appropriate for the future research. Our investigation has shown that due to the data scattering the simple linear term has not been a best choice for the modeling regardless of the apparent linear trend. Otherwise high order polynomials have caused model overfitting. We have discovered that quadric term is the most suitable for the relation modeling of such a kind of the data like temperature and energy parameters of the smart homes. Resulting regression graphs are shown in Fig. 2-4.

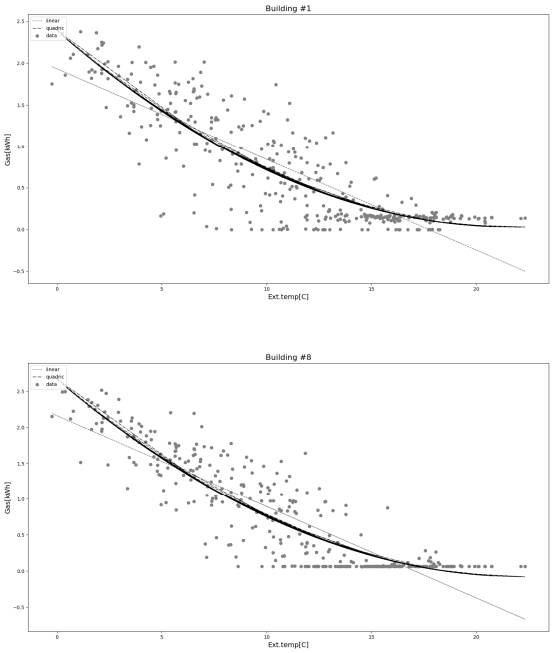

Fig. 2. Regression curves for the external temperatures and gas

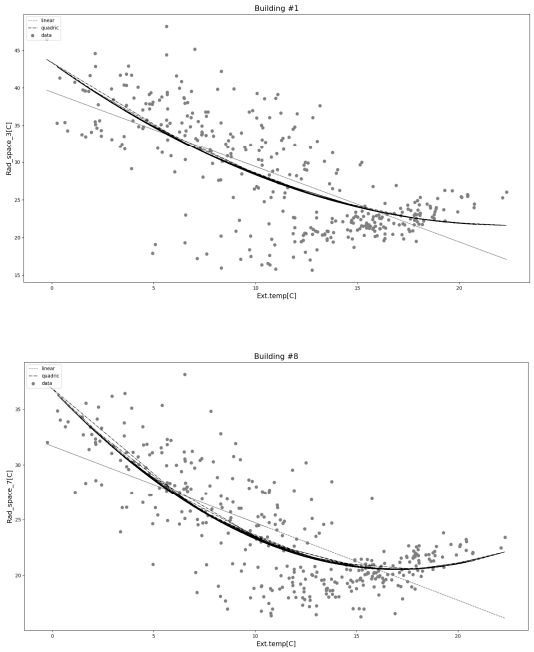

Fig. 3. Regression curves for the external temperatures and surfuce radiator temperatures

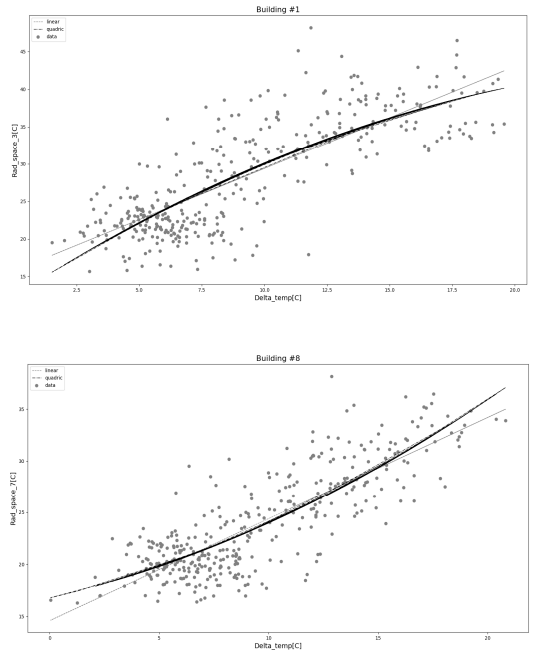

Fig. 4. Regression curves for the delta and surface radiator temperatures

In Fig. 2 the linear and quadric regression curves have been built to investigate relations between the external temperatures and the gas usage for two arbitrary chosen building #1 and building #8; in Fig. 3 these types of curves have been built for the external and surface radiator temperatures for two rooms (spaces) of the buildings #1, 4; in Fig. 4 we have used difference temperature $deltaT_i = T_i^{int} - T_i^{ext}$ between the

internal air temperatures T_i^{int} selected for the same room and external air temperature correspondingly to the time step.

IV. Discussion and conclusions

For each of the regression curves calculated with using equations (3)-(8) the following regression coefficient (c), mean squared errors (mse), r^2 values, Pearson's (Pr. corr.) and Spearman's (Spr. corr.) correlation coefficients have been obtained (Tables 1, 2):

TABLE 1. REGRESSION RESULTS FOR BUILDING #1

	Linear	Quadric	Pr. corr.	Spr. corr.
	c., mse, r^2	c., mse, r^2		
Gas/Ext.	(1.93, -0.1), 0.13, 0.7	(2.38, -0.22, 0.004), 0.11, 0.74	-0.84	-0.8
Ext/Rad	(39.4, -1), 26.7, 0.0	(43.3, -1.9, 0.042), 25, 0.52	-0.71	-0.65
Delta/Rad	(15.1, 1.36), 19.1, 0.64	(12.4, 2.14, -0.03), 18.7, 0.65	0.8	0.77

TABLE 2. REGRESSION RESULTS FOR BUILDING #8

	Linear	Quadric	Pr. Corr.	Spr. Corr.
	c., mse, r^2	c., mse, r^2		
Gas/Ext.	(2.16, -0.1), 0.13, 0.7	(2.66, -0.25, 0.006), 0.1, 0.81	-0.87	-0.86
Ext/Rad	(31.7, -0.7), 11.4, 0.53	(36.8, -1.91, 0.056), 9.02, 0.63	-0.72	-0.65
Delta/Rad	(14.6, 0.98), 7.85, 0.67	(16.8, 0.48, 0.02), 7.6, 0.7	0.82	0.76

As we can see from Tables 1, 2 quadric terms for the regressions are more appropriate for the modeling of the energy and thermal components of the smart home data. The mean squared errors and r^2 values point to the presence of the non-linear relations. Also, Pearson's and Spearman's correlation coefficients indicate strong relations between the data components with positive and negative values corresponding to the reverse and direct relations, respectively.

These results have confirmed a possibility of using such relations, i.e., regression models not only for the problem of the simple forecasting of energy data, but also for the development of forecasting neural networks combined with regression models, energy disaggregation problem, etc. Also, the calculation of the relations between smart home data components can be useful in the case of determination of heated/non heated spaces and preliminary estimation of thermophysical parameters (heat loss rate, cooling rate, heat capacity) of smart homes (inverse problem in the statistical and direct physical approaches).

Acknowledgment

We thank the research team from Loughborough University, UK: Dr. Steven Firth, Prof. Tarek Hassan, Dr. Tom Kane, Dr. Michael Coleman and PhD researcher Vanda Dimitrou for sharing open access sensor data of smart homes under REFIT project and answering our questions.

References

[1] Zehnder, M.; Wache, H.; Witschel, H-F.; Zanatta, D. and Rodriguez, "Energy saving in smart homes based on consumer behavior: A case study," 2015 IEEE First International Smart Cities Conference (ISC2), Guadalajara, pp. 1-6.

[2] J. W. Moon, M. H. Chung, H. Song, S.-Y. Lee, "Performance of a Predictive Model for Calculating Ascent Time to a Target Temperature," MDPI-Energies, vol. 9, pp. 1090, 2016.

[3] Yanni Zhai, Xiaodong Cheng, "Design of smart home remote monitoring system based on embedded system," 2011 IEEE 2nd International Conference on Computing, Control and Industrial Engineering, pp.41-44, 2011.

[4] D. Schweizer, M. Zehnder, H. Wache, H. Witschel, "Using consumer behavior data to reduce energy consumption in smart homes," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 1123–1130.

[5] D. B. Vilar, C. de Mattos Affonso, "Residential energy management system with photovoltaic generation using simulated annealing," 2016 13th International Conference on the European Energy Market (EEM), Porto, pp. 1-6.

[6] R. Hu, R. Skorupski, R. Entriken, Y. Ye, "A Mathematical Programming Formulation for Optimal Load Shifting of Electricity Demand for the Smart Grid for the Smart Grid," IEEE Transactions on Big Data, vol. 10, pp. 2332-7790, December 2016.

[7] Kanghang He, Lina Stankovic, Jing Liao, Vladimir Stankovic, "Non-Intrusive Load Disaggregation Using Graph Signal Processing," IEEE Transactions on Smart Grid, vol. 9, pp. 1739- 1747, 2016.

[8] T. Kane, S. Firth, T. Hassan, V. Dimitrou, "Heating behaviour in English homes: An assessment of indirect calculation methods," in Energy and Buildings, vol. 148, pp.89-105, 2017.