Renewable and Sustainable Energy Reviews 47 (2015) 332–343

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: www.elsevier.com/locate/rser

Regression analysis for prediction of residential energy consumption

Nelson Fumo*, M.A. Rafe Biswas

Mechanical Engineering Department, The University of Texas, Tyler, United States

ARTICLE INFO

Article history:
Received 14 July 2014
Received in revised form
9 January 2015
Accepted 8 March 2015
Available online 30 March 2015

Keywords:
Energy consumption
Residential buildings
Linear regression
Energy regression models

ABSTRACT

The considerable amount of energy consumption associated to the residential sector justifies and supports energy consumption modeling efforts. Among the three approaches to develop energy models, statistical approaches are a good option to avoid the burden associated to engineering approaches when observed/measured data is available. Among the statistical models, the linear regression analysis has shown promising results because of the reasonable accuracy and relatively simple implementation when compared to other methods. In this study, simple and multiple linear regression analysis along with a quadratic regression analysis were performed on hourly and daily data from a research house. The time interval of the observed data showed to be a relevant factor defining the quality of the model. Multiple linear regression models using the outdoor temperature and solar radiation offered improved coefficient of determination, but deteriorated root mean square error emphasizing the importance of using both parameters to assess and compare models. The content and structure of the paper has been devised to become a comprehensive material to be considered as the starting point for future work in this interesting research area. This paper also conveys the authors' belief that the future of residential energy forecasting is moving toward the development of individual models for each household due to the availability of data from smart meters, as well as the development of friendly and easy-to-use engineering software.

Contents

* Corresponding author.
E-mail address: nfumo@uttyler.edu (N. Fumo).

http://dx.doi.org/10.1016/j.rser.2015.03.035
1364-0321/

Nomenclature

Acronyms

CDD	cooling degree-days
CDA	conditional demand analysis
EIA	the U.S. Energy Information Administration
GHR	global horizontal radiation
GDP	gross domestic product
HDD	heating degree-days
HVAC	heating, ventilating, and air conditioning
OECD	organization for economic co-operation and development
PCA	principal component analysis
RECS	residential energy consumption survey
RMSE	root mean square error
SHEU	survey of household energy use
SSE	sum of square errors
SST	total sum of squares

Symbols

β	regression coefficient
$\hat{\beta}$	estimated regression coefficient
D	dummy variable
$E, \overline{E}$	electricity, and average
ε	error
GHR	global horizontal radiation
k	number of regression coefficients
n	number of observations (data)
τ	number of previous time periods
T	temperature
$X, \overline{X}$	predictor variable, and average
$Y, \overline{Y}$	response variable, and average
$\hat{Y}$	fitted or predicted variable

Subscripts

adj	adjusted
db	dry-bulb
HP	heat pump
i, j	variables index
p	number of predictor variables
t	time

1. Introduction

People spend the majority of their lifetime inside buildings, which implies a large demand of energy in order to satisfy occupational activities as well as thermal comfort. Fig. 1 shows the 2015 estimated total energy consumption (includes electricity related losses) associated to the residential end-use sector for the U.S. and countries of the Organization for Economic Co-operation and Development (OECD) and non-OECD countries [1]. Data is given on a relative basis to the world consumption, as well as the national consumption when compared to the other end-use sectors (commercial, industrial, transportation). At the national level, the U.S. residential sector consumes 21%, which is higher than the reference value of any of the groups (world, OECD, and non-OECD). On the other hand, from Fig. 1, it can also be noted that the residential sector of the U.S. represents 17% of the world's residential energy consumption.

Beyond the high contribution of the residential end-use sector on energy consumption, this sector 'is largely an undefined energy sink' when compared to the other three end-use sectors [2]. Swan and Ugursal [2] explain why the commercial, industrial, and transportation sectors are better understood, and point out that building characteristics, occupant behavior, privacy issues for collection and sharing data, and prohibitive cost of sub-metering are the factors that contribute to the 'undefined' cataloging of the residential sector. This condition of the residential sector and the substantial energy consumption of the sector in every country support efforts aiming the understanding of energy use for the sake of energy and emissions reduction.

In a global context, due to energy security, economics, and environmental concerns, governments as well as public and private sectors strive in the search of improving building energy performance. In this sense, the Building America program [3] offers resources on technology and research to reduce energy consumption and promote technology deployment. Besides, improvements to reduce grid energy consumption has been investigated for different alternative energy systems from which thermal storage systems may be an alternative to boost the efficiency of renewable energy technologies for space heating and cooling [4,5].

For new buildings, as well as for retrofits, the analysis of alternative designs can be assessed through simulations in order to determine the most efficient and cost-effective options. Approaches for building energy consumption modeling can be classified as statistical or black-box, hybrid or gray-box, and engineering or white-box [6]. The accuracy of an approach depends on the information that is available for the purpose of the approach. Statistical approaches need measured data, but not buildings characteristics, while the engineering approaches need building characteristics but not data, at least when a calibrated model is the goal. Among the statistical approaches, regression techniques deserve attention due to:

- Relatively ease to implement.
- Requirement of less computational power than other statistical approaches (genetic algorithms, neural networks, support vectors machine).
- Satisfactory prediction ability.
- Increased availability of data through smart metering.

As it may be noted from the previous paragraphs, this paper focuses on whole-building energy consumption or energy demand. However, there are other two areas of research related to the use of regression analysis to estimate residential energy consumption. The first area is regarding power demand. A review work by Grandjean et al. [7] covers the 'electric load curve models' which is the name given to the equations/models that can be used to predict domestic power demand. To have a better sense of the difference between both approaches, energy demand and power demand, Gradjean et al. stated 'The influence of the human behavior on the domestic power demand is so important that there is every chance for instance that two

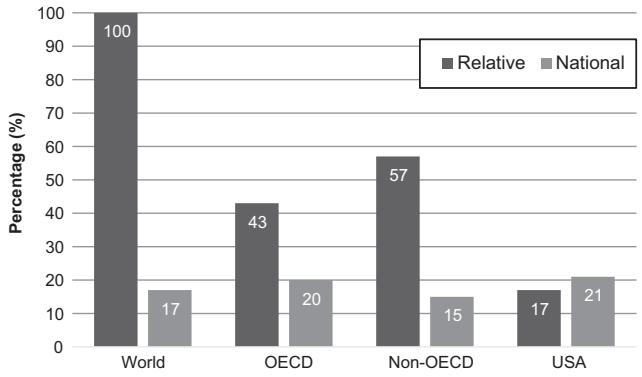

Fig. 1. 2015 – Residential end-use sector energy consumption.

households with the same daily energy consumption will not show a similar load curve.' The power demand approach is of interest to study topics mainly related to the need of predicting the peak power demand to analyze issues related to the electric network. The second area is regarding the energy demand at the sector scale. Kialashaki and Reisel [8] developed three multiple linear regression models using a stepwise linear regression algorithm to identify the relevant predictor variables in the models among the following predictor variables investigated: resident population, gross domestic product, household size, median household income, cost of residential electricity, cost of residential natural gas, and cost of residential heating oil. The data corresponds to the United States (U.S.) for the year 1984–2010. The sector energy demand approach is relevant to assist in planning for future energy needs, amount, and source. At this point, it is also important to mention that this paper does not cover regression methods for passive homes or multifamily buildings, for which the reader can refer to Refs. [9–12] [11,], correspondingly.

This paper is an effort to collect and summarize the most relevant information on regression analyses on prediction of energy consumption in residential buildings with the idea of provides a comprehensive reference for future studies. As a general outline, first a background of regression analysis is given; second, the basics on building energy regression models are treated; then, previous works on whole-building energy consumption prediction with regression analysis are summarized; and finally, a case study is presented to illustrate the application of linear regression analysis.

2. Basics of regression analysis

This section aims to give the reader an insight on statistical concepts related to regression analysis for better understanding of the literature review, and it is based mainly on references [13–17]. Multiple methods and regression algorithms are available for regression analysis, and specialized software, such as SPSS [18], SAS [19], SIMCA [20], STATISTICA [21], R [22], STATGRAPHICS [23] and NCSS [24] are used for applications. In building energy analysis the most common method used for regression analysis is the least squares method.

Regression analysis is a methodology that allows finding a functional relationship (model or equation) among *response* or *dependent* variables and *predictor*, *explanatory* or *independent* variables. When dealing only with one response variable, the regression analysis is called *univariate* regression; while when dealing with two or more response variables, the regression is called *multivariate* regression. For complex systems, such as the energy consumption in buildings, the regression analysis should be viewed as an iterative process; i.e. a process in which the outputs are used to diagnose, validate, criticize, and possibly modify the inputs.

2.1. Linear regression

The univariate linear regression analysis attempts to model the relationship among variables by fitting a linear equation to the data. When there is more than one predictor variable (multiple linear regression), the linear fitting is attempted by keeping constant all but one of the predictor variables. It should be understood that a relationship among a response variable and a predictor variable does not necessarily imply that the predictor variable causes the response variable, but that there is some significant association between the two variables. For example, the heat flux through the walls is a function of the thermal conductivity of walls, but also of the difference in temperature, which is the driving force of the heat flux. In other words, there is a significant association between the heat flux and the thermal conductivity of the walls, but the thermal conductivity does not cause the heat flux. There are two main types of regression analysis techniques that are used accordingly with the complexity of the relationship among variables, the simple linear regression and the multiple linear regression. In this section linear regression approaches are discussed based on the classification shown in Table 1, which also summarizes the equations associated to each approach.

2.1.1. Simple linear regression

The simple linear regression has an equation of the form

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

where Y is the response variable, X is the predictor variable, β_0 and β_1 are the *regression coefficients* or *regression parameters*, and ε is an error to account for the discrepancy between predicted data from Eq. (1) and the observed data. The predicted value form of Eq. (1) is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{2}$$

where $\hat{Y}$ is the *fitted* or *predicted* value and $\hat{\beta}$ are estimates of the regression coefficients.

The difference between fitted and predicted values is that the fitted value refers to the case where the values used for the predictor variable correspond to one on the n observations of the observed data used to find $\hat{\beta}$, but the predicted values are obtained for any set of values of the predictor variables different to the observed data.

2.1.2. Multiple linear regression

The multiple linear regression, or univariate multiple regression, is the generalization of the simple linear regression model. The model in multiple linear regression allows more than one predictor variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{3}$$

where Y is the response variable, X_1, X_2, $\ldots X_p$ are the predictor variables with p as the number of variables, β_0, β_1, $\ldots \beta_p$ are the regression coefficients, and ε is an error to account for the discrepancy between predicted data and the observed data. The predicted value form of Eq. (3) is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p \tag{4}$$

where $\hat{Y}$ is the *fitted* or *predicted* value and $\hat{\beta}$ are estimates of the regression coefficients.

As said for the simple linear regression, the difference between fitted and predicted values is that the fitted value refers to the case

where the values used for the predictor variables correspond to one on the n observations of the observed data, but the predicted values are obtained for any set of values of the predictor variables. In regression analysis, for estimation and testing purposes, $n > p+1$.

2.1.3. Quality of the model

The quality of fit of the linear model to a given set of observed data can be judged by using the *coefficient of determination* (R^2)

$$R^2 = [Cor(Y,\hat{Y})]^2 = 1 - \frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\bar{y})^2} \tag{5}$$

where $Cor(Y,\hat{Y})$ is the *correlation coefficient*, and the terms $\Sigma(y_i-\hat{y}_i)^2$ (or $\Sigma(Y-\hat{Y})^2$) and $\Sigma(y_i-\bar{y})^2$ (or $\Sigma(Y-\bar{Y})^2$) are called *sum of squared errors* (SSE) and *total sum of squares* (SST), respectively.

The value of R^2 varies between 0 and 1; a value of $R^2 = 0.9$ indicates that 90% of the total variability in the response variable is accounted for by the predictor variables. However, a large value of R^2 does not necessarily mean that the model fits the data well. Thus, a more detailed analysis is needed to ensure that the model can satisfactorily be used to describe the observed data and predict the response for another set of data different from the one used to generate the model.

For a multiple linear regression model, the adjusted coefficient of determination is defined in terms of R^2 as

$$R^2_{adj} = 1 - (1-R^2)\frac{n-1}{n-p-1} \tag{6}$$

where k is the number of regression coefficients. Based on Eqs. (2) and (4), it can be noted that $k = p+1$. In Eq. (6), the term $n-p-1$ is used to maintain consistency with format found in the literature, thus it can also be noted that $n-p-1 = n-k$.

The root mean square error (RMSE) of the model is another parameter to measure the quality of the fitting, which is a measure of the scatter in the data around the model. The RMSE for a simple linear model is computed as

$$RMSE = \sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n}} = \sqrt{\Sigma\frac{(Y-\hat{Y})^2}{n}} \tag{7}$$

and for a multiple linear model is computed as

$$RMSE = \sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n-k}} = \sqrt{\Sigma\frac{(Y-\hat{Y})^2}{n-k}} \tag{8}$$

2.1.4. Multivariate linear regression

In Section 2.1.2 it was stated that multiple linear regression is the generalization of the simple linear regression technique. Similarly, the multivariate linear regression technique, or multivariate multiple linear regression technique, can be seen as a generalization of the multiple linear regression technique. The multivariate linear regression analysis attempts to interpret possible linear relationships among multiple response variables and multiple predictor variables. Eq. (4) can be rewritten as Eq. (9) to illustrate that two or more response variables are being analyzed

$$\hat{Y}_i = \hat{\beta}_{i,0} + \hat{\beta}_{i,1}X_1 + \hat{\beta}_{i,2}X_2 + \cdots + \hat{\beta}_{i,p}X_p \tag{9}$$

In multivariate regression, not only the predictor variables are correlated with each other, but the response variables are correlated to each other and, of course, with the predictor variables. As an example, let us assume that we are interested in finding the relationship between the electricity consumption of a household and the environmental variables including outdoor temperature, outdoor relative humidity, and solar radiation. These three predictor variables are correlated with each other. If the electricity consumption is desegregated in HVAC equipment, lighting, and appliances, these response variables are correlated to each other.

2.1.5. Linear regression in matrix form

The general matrix form of Eq. (3) ($Y = X\beta + \varepsilon$) is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

It can be noted that the format of the vector of regression coefficients is different from the response vector and error vector to illustrate that $n > p+1$.

The least-squares regression method estimates the regression coefficients of the model that minimize the sum of the square error between the predicted and actual observations from a given set of data. In a matrix form, the estimated regression coefficients can be obtained using the transpose ($'$) and inverse ($^{-1}$) matrix functions as follows:

$$\hat{\beta} = (X'X)^{-1}(X'Y) \tag{10}$$

and the estimated values can be obtained by

$$\hat{Y} = X\hat{\beta} \tag{11}$$

If there is a collinearity (see Section 2.3) among predictor variables, the square matrix $(X'X)$ is singular and does not have an inverse.

Eqs. (10) and (11) also applies for multivariate regression. For a data set containing 2 response variables and 3 predictor variables, the matrix format of a multivariate form of Eq. (3) is

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \\ \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} \end{bmatrix}$$

The use of Eq. (9) allows one to find the matrix of estimated regression coefficients $\hat{\beta}$. However, based on what was stated in Section 2.1.4, the problem can also be solved for each response variable (Y_1, Y_2) separately in a multiple regression approach. In addition, the vectors $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained with the same regression coefficients as in the matrix $\hat{\beta}$ can be obtained with the multivariate approach.

2.2. Nonlinear regression

Like linear regression analysis, the nonlinear regression can be single or multiple variable regression. For simplicity, the discussion is based on one predictor variable. On the other hand, with the background gained in previous sections, it will be noted that some explanatory ideas are also applied to the linear regression analysis.

In general, the form of the function (f) describing the relationship between the response variable and the predictor variable is not known, but it can be formulated as

$$Y = f(X,\beta) \tag{12}$$

with a set of k unknown parameters (regression coefficients) $\beta = (\beta_1, \ldots, \beta_k)$.

For a set of observed data with pairs (y_1,x_1), …, (y_n,x_n) with no error associated to the response or predictor, Eq. (12) applies as the ideal case. However, since measurement errors will make the pairs to be out of the fitting function, at the best case Eq. (11) will be correct only on average. Therefore, Eq. (12) can be expressed as

a mean function by

$$\hat{Y} = f(X, \hat{\beta}) \quad (13)$$

Since the regression analysis only allows to find the unknown parameters and not the form of the function describing the relationship between the response and predictor variables, based on a graphical representation of the observed data and/or experience of the analyst, a defined function should be used. Specialized software come with known models available from a library, but also allow to create a new formula (model). The following are some examples of models to give the reader an idea of nonlinear models:

- Asymptotic regression (from SPSS library) $f(X, (\beta_1, \beta_2, \beta_3)) = \beta_1 + \beta_2\, e^{X\beta_3}$.
- Michaelis Menten (from SPSS library) $f(X, (\beta_1, \beta_2)) = (\beta_1\, X/X + \beta_2)$.
- $f(X, Z, (\beta_1, \beta_2, \beta_3)) = \beta_1/(1 + \beta_2(X + \beta_3 Z))$ (from example in Ref. [13]).

2.3. Collinearity

Collinearity or multicollinearity refers to the existence of strong linear relationships among the predictor variables, which means that one predictor variable can be near-linearly predicted from the others. When there is no linear relationship among predictor variables at all, they are said to be orthogonal. The lack of orthogonality among the predictor variables usually is not strong enough to affect the analysis or the ability of the entire number of predictors to predict the response variables. In other words, the lack of orthogonality does not diminish the usefulness of the model, at least within the sample data used to find the regression coefficients. However, this condition can produce ambiguous results that are associated with unstable estimated regression coefficients and affects the calculations associated to individual predictors. Instability in the estimated coefficients can be indicated by large changes in the estimated regression coefficients when a variable is added or deleted, or when a data point is altered or dropped. When dealing with collinearity, the principal component analysis (PCA) method is one of the most common ways to reduce collinearity [25]. The PCA, contrary to grouping methods such as cluster analysis, is a one-sample technique that uses orthogonal transformation to obtain a set of values of linearly uncorrelated variables called principal components, which can be equal to or less than the original number of predictor variables.

3. Basics on energy regression models

3.1. Top-down and bottom-up approaches

Top-down approaches [2,6] use information on a high hierarchical level of energy consumption which implies the consideration of variables defining the overall energy consumption instead of individual end-uses. Top-down approaches identify factors defining changes in energy consumption trends on the long-term within the residential sector which are used to identify future energy demand and improving efficiency and building design. Top-down approaches can also be used to identify benchmarks for end-uses. As an example, Tso and Guan [26] performed a regression analysis to understand effects of environmental indicators and household features on residential energy consumption using three sources of data: the 2009 Residential Energy Consumption Survey (RECS) micro dataset, the 2009 Annual Energy Review published by the U.S. Energy Information Administration (EIA), and the macroeconomic statistics published by Bureau of Economic Analysis. An example of the use of information from top-down approaches is the Home Energy Yardstick tool [27]. This tool uses statistical algorithms to take into account the effects of local weather, home size, and number of occupants, and compares the energy consumption to generate a score based on data obtained from the RECS. On the other hand, the RECS End-Use Models [28] is an example of top-down approach at the end-use level. The RECS end-uses model 'is a set of equations designed to disaggregate a RECS sample household's total annual fuel consumption into end uses such as space heating, air conditioning, water heating, refrigeration, and so on.' These models are needed because the cost of using submitters on sample households is prohibitively expensive.

Bottom-up approaches [2,6] aims to characterize a house archetype or houses in a geographic region by extrapolating energy consumption obtained from models defining the energy consumption at the end-use or house level. For better accuracy small samples with similar characteristics can be used to characterize the segment of the residential sector of interest. Bottom-up approaches can be statistical, hybrid, or engineering. From the statistical approaches, models using regression analysis is the focus of this paper.

3.2. Energy signatures

An energy signature [29–33], also called thermal performance line or building energy performance line or heat balance equation, is the best fit correlating energy consumption with weather variables. An energy signature is based on a physical model that is adapted or modified to meet the format needed for a regression method. The approach of obtaining the energy signatures is a system identification approach, also known as inverse modeling. Since an energy signature is a representation of the actual energy performance of a building, measured energy consumption and climate data is needed to perform the regression. The most common method used for regression of energy signatures is the least squares method. Energy consumption can be the total of the building or can correspond to a systems of interest such as cooling, heating, hot water, etc. Outdoor temperature is the natural weather variable used, but solar radiation and wind effects could also be incorporated into the regression equation. However, models using heating and cooling degree days (HDD and CDD) are also common.

Energy signatures are commonly derived as static models for which the time period (time-step) are chosen long enough in order that the heat stored or release from the building is very small when compared to the total energy for the same time interval. Hammarsten [29] stated that for a static model the time period should not be shorter than an hour. However, the time period normally will depend on the available data. Although energy data can be obtained for time periods smaller than the readily available monthly energy consumption from utility bills, weather data is more difficult to obtain. When data of one of the variables has been recorded for a smaller time period than the one of analysis, the recorded data should be added if the data corresponds to an energy consumption variable or should be averaged if the data corresponds to a weather variable. General information on weather data can be found in [6], while temperature and HDD/CDD, for example, can be obtained from [34] and [35], respectively.

If the time period of analysis is an hour or lower, the model must be developed as a dynamic model to account for the thermal inertia of the building to avoid inaccuracy due to the lag between the energy consumption and the weather variable [29,33]. Dynamic models can be represented by a general autoregressive

model as

$$\hat{Y}_t = \hat{\beta}_0 + \sum_{j=0}^{\tau} \hat{\beta}_{1,j} X_1(t-j) + \ldots + \sum_{j=0}^{\tau} \hat{\beta}_{p,j} X_p(t-j) + \sum_{j=1}^{\tau} \hat{a}_j Y(t-j) \quad (14)$$

where τ is the number of previous time periods (for example an hour) to be included in the analysis, $\hat{\beta}_{1,j}, \hat{\beta}_{2,j}, \ldots, \hat{\beta}_{p,j}$ are the regression coefficients associated to the respective predictor variables ($X_1, X_2, \ldots, X_p$) for the time of analysis and for previous times ($t-j$) and $\hat{a}_j$ are regression coefficients associated to the lag terms of the response variable. As it can be seen from Eq. (14), autoregressive means that the response variable depends linearly on its immediate past values.

3.3. Conditional demand analysis

Conditional demand analysis (CDA) is a method for disaggregating the total household electric load into the individual components associated with a particular end-use or appliance. The method does not require direct observation on specific end-use energy usage, but once the electric load curves or demand functions are found, they can be used to estimate the monthly and annual average energy consumption associated to each of the end-uses considered. As referenced by Bartels and Fiebig [36] and Aigner et al. [37], this regression-based approach was pioneered by Parti and Parti [38]. The equation proposed by Parti and Parti to estimate the total electricity consumption is

$$E = \sum_{i=0}^{N} \overline{E}_i [D_i] + \sum_{i=0}^{N} \sum_{j=1}^{p} \hat{\beta}_{ij} \left[(X_j - \overline{X}_{ij}) D_i \right] \quad (15)$$

with

$$\overline{E}_i = \hat{\beta}_{i0} + \sum_{j=0}^{p} \hat{\beta}_{ij} (\overline{X}_{ij}) \quad (16)$$

where N is the number of end-uses or appliances considered in the model, $\overline{E}_i$ is the average energy used through the ith end-use, D_i is a dummy variable that takes the value of one for those households possessing the ith end-use and zero otherwise, p is the number of predictor variables, X_j is a vector of predictor variables defining the ith end-use, $\overline{X}_{ij}$ are the average values of the predictor variables in households that possess the ith end-use, and $\hat{\beta}_{ij}$ are the regression coefficients that capture the impact of the jth predictor variable on the ith end-use. For $i=0$ ($\overline{E}_0$), there is no appliance specified and it accounts for energy consumed by unspecified end-uses or appliances. In their study [38], regression coefficients were obtained from 12 consecutive monthly regression analysis for disaggregation of the total household demand into 16 specified end-use categories and one unspecified category. Results of the CDA was used as part of an econometric analysis to estimate price and income elasticity based on data from 5286 households in San Diego County, CA.

As used by Aigner et al. [37], CDA can be done for time periods of an hour. In their study [37], regression coefficients were found for each hour of a 24-h period involving the 14 predictor variables proposed to estimate average hourly energy consumption over the days of the month. In their conclusion, the method is considered to offer a significant lower-cost alternative to direct metering of most of the major appliances. However, they point out that although results showed well-defined load shapes for many appliances, the load levels often seemed questionable with justification on the simple model specification they used.

As an example of the use of the CDA in a top-down approach, Lafrance and Perron [39] presented conclusions from a temporal analysis of three large scale surveys for the Québec (Canada) residential sector. The end-uses investigated were space heating, space cooling, water heating, and appliances, with the following predictor variables used in combinations, according to the end-use: degree-days, space area, number of persons, wood cord purchase, income, size of the water heater, teenage rate, and dishwasher and pool bath possessions. Their conclusions illustrate how the CDA allows identification of changes in the patterns of energy use by end-uses through the long periods and to visualize efficiency comparison through an analysis of system types.

In an effort to compare the CDA with other approaches, Aydinalp-Koksal and Ugursal [40] compared the CDA with neural network and engineering approaches in a study based on the 1993 Survey of Household Energy Use (SHEU-1993). The SHEU-1993 collected detailed data on the energy consumption habits of households in Canada. The authors concluded that the CDA model used was capable of accurately predicting the energy consumption in the residential sector as well as the other two models used.

Table 1
Classification of linear regression approaches.

Type of regression		Response variables	Predictor variables	Regression equation	Quality of model
Univariate	Simple	1	1	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ (2)	$R^2 = 1 - \frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \overline{y})^2}$ (5) $RMSE = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{n}}$ (7)
	Multiple	1	≥ 2	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$ (4)	$R^2_{adj} = 1 - (1 - R^2) \frac{n-1}{n-p-1}$ (6)
Multivariate		≥ 2	≥ 1	$\hat{Y}_i = \hat{\beta}_{i,0} + \hat{\beta}_{i,1} X_1 + \hat{\beta}_{i,2} X_2 + \cdots + \hat{\beta}_{i,p} X_p$ (9)	$RMSE = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{n-k}}$ (8)

Table 2
Especial application of regression approaches.

Approach	Response variables	Predictor variables	Regression equation	Quality of model
Autoregressive	1	≥ 2	$\hat{Y}_t = \hat{\beta}_0 + \sum_{j=0}^{\tau} \hat{\beta}_{1,j} X_1(t-j) + \ldots + \sum_{j=0}^{\tau} \hat{\beta}_{p,j} X_p(t-j) + \sum_{j=1}^{\tau} \hat{a}_j Y(t-j)$ (14)	$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum} (y_i - \overline{y})^2$ (5) $RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$ (7)
CDA	1	≥ 2	$E = \sum_{i=0}^{N} \overline{E}_i [D_i] + \sum_{i=0}^{N} \sum_{j=1}^{p} \hat{\beta}_{ij} [(X_j - \overline{X}_{ij}) D_i]$ (15) $\overline{E}_i = \hat{\beta}_{i0} + \sum_{j=0}^{p} \hat{\beta}_{ij} (\overline{X}_{ij})$ (16)	$R^2_{adj} = 1 - (1 - R^2)_{\frac{n-1}{n-p-1}}$ (6) $RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k}}$ (8)

However, they pointed out that the CDA model was unable to evaluate the effects of some socio-economic factors due to statistical considerations. This is explained by the fact that some variables could not be included in the model due to their correlation, which produces collinearity. Another limitation found for the CDA model is that the model is not flexible in evaluating end-uses and energy saving scenarios.

On the other hand, Newshamn and Donnelly [41] suggest that CDA can be used to estimate energy consumption as a means to identify incentives and policies most likely to be cost-effective. They applied CDA to a raw data of 9773 Canadian households to estimate the average annual energy use of various electrical and natural gas appliances. The regression model for electricity accounts for 25 predictor variables with adjusted coefficient of determination of 0.525. While the regression model for natural gas accounts for 14 predictor variables with adjusted coefficient of determination of 0.792. The model was used to analyze cases for energy reductions associated with certain appliance upgrades and behaviors to support the idea of using this type of results for incentives and policies regarding residential energy consumption.

The autoregressive and CDA approaches discussed are special applications of regression approaches in the analysis of energy consumption. Like the comparison of the approaches are summarized in Tables 1 and 2 summarizes these two approaches.

4. Studies on residential regression analysis

Westergren et al. [33] present results of regression models to estimate the heating energy consumption per unit of time (hour). Two static models and one dynamic model were evaluated and compared. The first static model is a simple linear model of the form of Eq. (2) with the predictor variable defined as the weekly average of the difference of indoor and outdoor temperatures. The second static model is a multiple linear regression model of the form of Eq. (4) that incorporates global horizontal solar radiation as the second predictor variable. The dynamic model is an autoregressive model with 11 free parameters accounting for hourly temperature difference and solar radiation with lag of two previous hours to the time of analysis. Energy consumption data used by Westergren et al. was obtained from hourly measurements on four houses and weather data from close climate station. Comparison shows similar results for the static and dynamic models with error of the estimations in the order of 2.5% and 9.0% depending on the size of the sample, the observation period, and the model used.

Raffio et al. [42] describe a four-step method to analyze energy performance of residential buildings with the intention of detecting building opportunities for energy improvement. The method can help to identify if a building is a good candidate for hot-water heater retrofit, programmable thermostat installation, envelope improvement or high-efficiency HVAC equipment retrofit. In the first step of the method, energy signature models are obtained by regressing the energy consumption from utility bills to actual average daily temperatures. Two regression equations of the form of Eq. (2) were obtained, one for gas usage and the other for electricity usage. The predictor variable is defined as the difference between the outdoor temperature and the balance point temperature[1]. If the outdoor temperature is lower than the balance point temperature, the house needs to be heated and the regression equation for gas usage applies. Contrarily, if the outdoor temperature is higher than the balance point temperature, the house needs to be cooled and the regression equation for electricity applies. The first regression coefficient (β_0) is a measure of the base load or load independent of weather conditions (outdoor temperature is equal to the balance point temperature), and the second regression coefficient (β_1) is a measure of the impact of the outdoor temperature on the energy consumption (the balance point temperature is constant). Accuracy of the method was satisfactory for the sample of houses used in the investigation, allowing identification of high hot water temperature setpoints, low efficiency hot water heaters, no nighttime set-backs, high rate of infiltration, and low furnace efficiency.

Catalina et al. [43] tested numerous models to find the best fit between the heating demand of single-family residences as the response variable and four predictor variables: shape factor, envelope *U*-value, window to floor area ratio, building time constant, and climate coefficient. The shape factor is defined as the ratio between the heated volume of the building and the total surface area in contact with the exterior; while the climate coefficient is computed as the difference between the heating set-point temperature and the average sol-air temperature (calculated using the monthly outdoor dry-bulb temperature, the monthly average global horizontal radiation, and a default exterior convection coefficient). The best model was found to be a polynomial, which predicted the synthetic[2] data used within 1.2–5.2% for the 270 validation scenarios. The synthetic data was obtained with hourly time-step simulations performed using the building simulation software TRNSYS®. The analysis of results showed a strong relationship between the shape of a building and the energy consumption, and that building thermal inertia has a significant impact on the energy demand.

Soldo et al. [44] tested the impact of solar radiation on models to predict natural gas consumption for heating space in a model house. Models tested include an auto-regressive model as a linear model, and a neural network and a support vector machine as nonlinear models. The model house has a heated space of approximately 100 m^2. To minimize human influence on gas consumption, the model house has only one equipment consuming natural gas, a boiler with central heating system. Natural gas consumption was recorded every hour, but data were resampled into a daily resolution. The prediction problem was formulated based on the dependence of the prediction value on past natural gas consumption, and past, current, and future predictor variable values. Results show that the use of solar radiation improves the accuracy of natural gas forecasting models. Comparison of linear and nonlinear models shows that although the nonlinear models have smaller training errors, these models do not improve the generalization ability on test data since the testing errors of the nonlinear models are slightly higher compared to the results obtained by linear models.

Elsawaf et al. [45] used regression analysis to evaluate the effectiveness of using heat pumps for space heating in four cities in eastern North Carolina (USA). The evaluation was based on comparison of energy consumption for homes with different methods of heating: heat pump, gas heating, electrical heater, and combination. The data used included information from surveys and actual energy consumption (gas and electrical). The multivariable linear regression analysis using the least squares method was performed with seven predictor variables: house size, number of occupants, number of stories, years since construction, house orientation, heating temperature, and a dummy variable (value of 0 or 1) to indicate if the home relies on heat pumps. Results supported the hypothesis that residential homes using heat pumps are more energy efficient and achieve higher energy savings than the other heating methods investigated.

Min et al. [46] presented results from a regression-based statistical analysis for modeling residential energy consumption in the United

[1] Balance point temperature: is that value of the outdoor temperature at which, for the specified value of the interior temperature, the building does not require heating or cooling.

[2] Synthetic data: data that is not obtained by direct measurement.

States (U.S.) at the zip-code level. Models were developed using data obtained from the U.S. Energy Information Administration on the Residential Energy Consumption Survey, and the models were tested for prediction using data from the U.S Census 2000. The ordinary least squares method was used to find linear and log-linear models for heating, cooling, water heating, and appliances as the response variables, with predictor variables including energy price, household characteristics, housing unit characteristics, regional fixed effects, and heating/cooling degree-days. The variable appliances included all other energy consumption not included in heating, cooling, or water heating. The total energy consumption of the variables heating, water heating, and appliances were computed as the aggregation of energy consumption from all the fuel types: electricity, natural gas, fuel oil, and liquefied petroleum gas; while electricity is the only fuel considered for cooling. The adjusted coefficient of determination for the linear/log-linear models of heating, cooling, water heating, and appliances were 0.594/0.825, 0.490/0.703, 0.295/0.343, and 0.409/0.518, respectively.

Chen et al. [47] performed correlation analysis among household variables and energy consumption, and used a multivariate regression analysis to explore the overall effect of socio-economic and behavioral variables on residential energy consumption of conditioned space in China. The data was obtained from surveys applied to 642 households during winter and 838 household during summer in the city of Hangzhou. Based on the results obtained, they point out: (1) socio-economic parameters explain 26.3% of variation in energy consumption of the conditioned space, which increases up to 28.8% when behavioral variables are included, (2) income accounts for 18% of variation in air conditioner, and (3) floor area accounts for 44% of variation in air conditioner.

Schleich et al. [48] used an ordinary least-squared regression analysis to investigate the impact of providing feedback on energy consumption. Data was obtained from 1500 households in Linz, Austria. In the study about half of the households received feedback on information about electricity saving measures, the pilot group, while the rest served as a control group. The households of the pilot group were able to choose between receiving feedback through a web portal or by a postal mail. In the regression analysis dummy variables were used to identify if the household received feedback from one of the two modes of feedback. The regression results suggested that the feedback led to an average of 4.5% of energy savings per household.

Gans et al. [49] investigated the effect of real-time usage information on residential electricity consumption in Northern Ireland. The data used corresponds to data from the Continuous Household Survey of Northern Ireland which is conducted year-round with approximately 300 household surveyed per month. The survey asks for information about dwelling, health, education, employment and welfare payments. To estimate the electricity demand, a regression model is used accounting for electricity price, household income, and variables influencing the electricity consumption such as weather, characteristics of the home and of the household, type of heating and appliances used. The model has dummy variables to consider the month of year when the household was interviewed. The regression coefficients were found using an ordinary least-squared method. Although the results do not document how households managed to reduce usage, usage reductions in the order of 10–17% were reported.

Mastrucci et al. [50] developed a multiple linear regression model to estimate natural gas and electricity consumption of single dwellings in the city of Rotterdam, Netherlands, using a bottom-up statistical approach. Dwellings characteristics such as type of dwelling, year of construction, floor surface, and number of occupants were used along with corresponding household records of natural gas and electricity consumption. The model was mainly developed for evaluation of typical refurbishment measures. Using the model it was estimated that energy saving potential are in the range of 41–68% and 5–12% for dwellings built before 1964 and between 1992 and 2005, respectively. For dwellings built after 2005, the potential for saving is estimated to be null. For evaluation of the quality of the model the coefficient of determination and the mean square error were used. The magnitude of the coefficient of determination were 0.718 and 0.817 for the natural gas and electricity models.

Nie and Kemp [51] investigated the increase of energy consumption in China during the period of 2002–2010 using data from China statistical yearbook 2011 and China energy statistical yearbook 2003–2011. They investigated the impact of changes in appliances, floor space, population, and energy source mix. The main factor contributing to energy consumption increase was appliances, followed by floor space per capita, with the energy mix the less important, and population the most stable. In order to predict electricity use, the authors developed regression based predictions to investigate the relationship between electricity consumption and ownership of appliances. The main conclusion was that the consumption of electricity will continue to rise despite a gradual saturation of demand.

Bianco et al. [52] investigated the residential and non-residential annual electricity consumption in Italy during the period 1970–2007 to develop a simple model to forecast electricity consumption. Simple and multiple regression models were developed using historical electricity consumption, gross domestic product (GDP), GDP per capita, and population. Results shows that the selected predictor variables are strongly correlated to the electricity consumption. For the residential models, the coefficient of determination for the simple and multiple regression models were found to be 0.975 and 0.990, respectively.

Ndiaye and Gabriel [53] performed a conditional demand analysis of 59 predictor variables to obtain a regression model with only 9 predictor variables that gave a coefficient of determination of 0.784. The model was developed to predict the electricity consumption of housing units in Oshawa (Ontario, Canada). The 9 predictors are number of occupants, house status (owned or rented), average number of weeks of vacation taken away from the house each year, type of fuel for the pool heater, type of fuel for the heating system, type of fuel for the domestic hot water heater, availability and type of an air conditioning system, and number of air changes per hour at 50 Pa. The data was collected based on three methods: survey, site audits, and audit of smart meters information.

Filippína et al. [54] using stepwise selection and multivariate analysis evaluate historical consumption of natural gas for heating in multifamily buildings. Data from 72 apartments belonging to different buildings, with different orientations and levels were used to generate 14 energy classes (clusters). The stepwise method applied by the authors allowed them to select the useful area, volume, envelope area, envelope area/useful area, envelope area/volume, passive zone area, volumetric heat loss coefficient, and weighted mean internal surface temperature as the variables categorizing the annual energy consumption.

5. Future of prediction of residential energy consumption

It is the authors' opinion that the future of residential energy consumption prediction will move toward to the analysis of individual dwellings, leaving behind models that use data from a sample to describe a population. This tendency will increase the accuracy of the predictions independently if a statistical or an engineering approach is used. The authors' opinion is based on the continuous effort on deployment of smart meters and the development of engineering

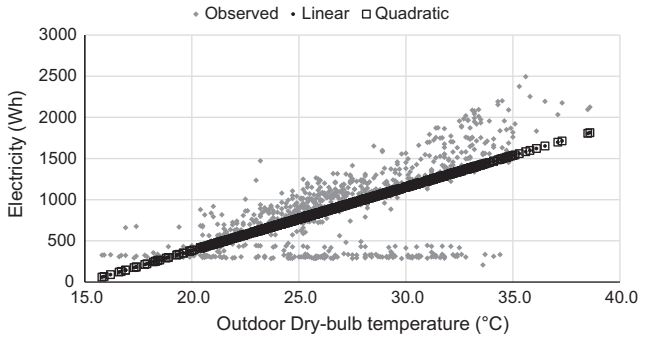

Fig. 2. Hourly total electricity as a function of outdoor dry-bulb temperature for simple linear regression.

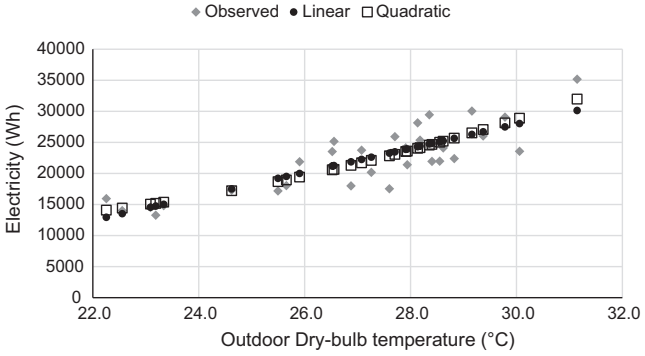

Fig. 3. Daily total electricity as a function of outdoor dry-bulb temperature for simple linear regression.

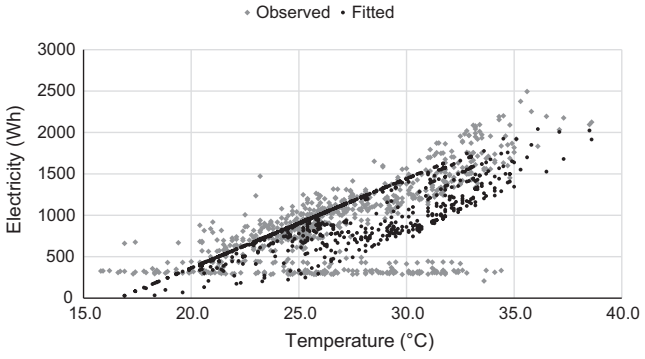

Fig. 4. Hourly total electricity as a function of outdoor dry-bulb temperature for multiple linear regression.

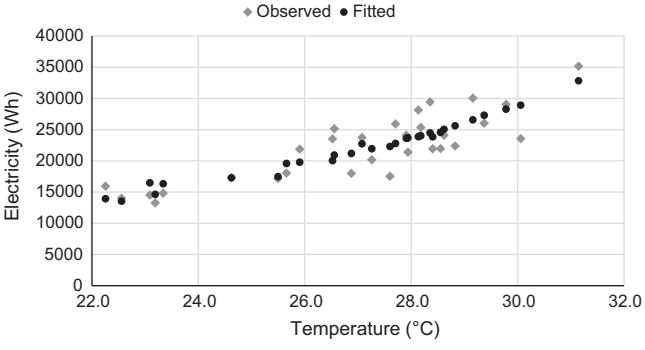

Fig. 5. Daily total electricity as a function of outdoor dry-bulb temperature for multiple regression.

software with a simple and friendly user interface. Regarding smart meters, the Green Button program [55] offers electronic access to energy consumption of individual dwellings at a resolution of 15 min. These data can be used to develop and validate statistical models, such as the regression models considered in this paper. Regarding engineering models, the National Renewable Energy Laboratory, in a continuous effort, has developed the software BEopt [56]. BEopt is a no-cost software with a friendly user interface capable of evaluation of residential building designs and identification of cost-optimal efficiency packages at various levels of whole-house energy analysis. The data from smart meters will make it possible to calibrate models developed with BEopt in an automatic manner using the Autotune methodology [57]. However, since engineering software requires inputs on the characteristics of the dwelling that it may need the participation of qualified personnel for accuracy. Therefore, the authors favor statistical models because they can be developed without the need of input parameters that describe the building characteristics. In order to illustrate the application of linear regression approaches as treated in Section 2, a case study to estimate the total energy consumption is given in the following section.

6. Case study: TxAIRE research house

The data used for the regression analysis presented in this section corresponds to the energy consumption and weather parameters recorded at the TxAIRE Research and Demonstration House #1 [58] during the month of June 2013. The house is unoccupied and all activities in the house are of an academic nature. The house uses only electricity and energy consumption is recorded every 5 min for the total electricity and HVAC equipment. Weather data is recorded also every 5 min by a weather station located at the research site. For this analysis, the data is compiled to obtain the hourly and daily data for the total energy consumption and weather parameters. The total (house) energy consumption (electricity) is computed in Wh associated to the time period of analysis (hourly or daily), while the outdoor dry-bulb temperature is computed in °C and the global horizontal radiation is computed in Wh/m^2 associated to the time period of analysis (hourly or daily). As a reference, the daily data is given in Appendix A.

6.1. Simple linear regression

For a simple linear regression model of the form of Eq. (2), the total electricity is the response variable and the outdoor dry-bulb temperature is the predictor variable. Using the matrix form of Eq. (9) in a MathCad® code or the Regression function on the Excel® Data Analysis package, the model for the hourly energy consumption is found to be

$$E_{hour} = -1156.2 + 77.0 T_{db} \quad (17)$$

with parameters of quality analysis as $R^2 = 0.423$ and $RMSE = 378$ Wh.

While the daily energy consumption is found to be

$$E_{day} = -30098.4 + 1933.7\, T_{db} \quad (18)$$

with parameters of quality analysis as $R^2 = 0.711$ and $RMSE = 2834$ Wh.

Since HVAC systems are responsible for a large portion of the total energy consumption in buildings, and because performance curves of HVAC system are not modeled based on linear models, but second or higher order polynomials, a second order regression analysis is performed to investigate the improvement of the model. Using the same tool (MathCad® or Excel®), the model for the hourly energy consumption is found to be

$$E_{hour} = 1325.2 - 111.5\, T_{db} + 3.5\, T_{db}^2 \quad (19)$$

with parameters of quality analysis as $R^2 = 0.444$ and $RMSE = 371\ Wh$.

Table 3
Summary of results for the case study.

Type of model	Interval	Equation	Predictor	R^2 (R^2_{adj})	RMSE	Electricity (Wh)		
						Min	Max	Avg
Simple linear	Hour	(17)	T_{db}	0.423	378	208	2495	922
Simple quadratic		(19)	T_{db}	0.444	371			
Multiple linear		(21)	T_{db} GHR	0.579	379			
Simple linear	Day	(18)	T_{db}	0.711	2834	13268	35165	22119
Simple quadratic		(20)	T_{db}	0.693	2920			
Multiple linear		(22)	T_{db} GHR	0.740	2920			

Table 4
Comparison of regression models and BEopt model.

Type of model	Interval	Equation	Predictor	R^2 (R^2_{adj})	RMSE
Simple linear	Hour	(17)	T_{db}	0.512	348
Simple quadratic		(19)	T_{db}	0.541	337
Multiple linear		(21)	T_{db} GHR	0.549	348
Simple linear	Day	(18)	T_{db}	0.701	2885
Simple quadratic		(20)	T_{db}	0.677	2995
Multiple linear		(22)	T_{db} GHR	0.705	2995
BEopt	Hour	n/a	n/a	0.315	412
	Day			0.691	2934

While the daily energy consumption is found to be

$$E_{day} = 42513.7 - 3626.7\,T_{db} + 105.6\,T_{db}^2 \quad (20)$$

with parameters of quality analysis as $R^2_{adj} = 0.693$ and $RMSE =$ 2920 Wh.

The observed (measured) hourly data, along with the fitted results from the linear and quadratic regressions, Eqs. (17) and (19), are shown in Fig. 2. While the results for the daily linear and quadratic regressions, Eqs. (18) and (20), are shown in Fig. 3.

6.2. Multiple linear regression

For a multiple linear regression model of the form of Eq. (4), the total electricity is the response variable and the outdoor dry-bulb temperature and the global horizontal radiation (GHR) are the predictor variables. Using the matrix form of Eq. (9) in a MathCad® code or the Regression function on the Excel Data Analysis package, the model for the hourly energy consumption is found to be

$$E_{hour} = -1795.6 + 107.0\,T_{db} - 0.67\,GHR \quad (21)$$

with parameters of quality analysis as $R^2_{adj} = 0.579$ and $RMSE =$ 379 Wh.

While the daily energy consumption is found to be

$$E_{day} = -30239.5 + 1882.9\,T_{db} + 0.213\,GHR \quad (22)$$

with parameters of quality analysis as $R^2 = 0.740$ and $RMSE =$ 2920 Wh.

The observed (measured) hourly data, along with the fitted results for the hourly and daily electricity consumption are shown in Figs. 4 and 5, respectively.

Table 3 summarizes the results from the case study.

6.3. Discussion

From Table 3, it can be noted that models based on a daily time interval offers better quality parameters, regardless of the type of model. This can be explained because as the resolution in time increases, the operation schedule of equipment and systems are more notorious; which is reduced by the averaging that results from increasing the resolution. To verify this statement, simple linear regression analysis were done for time intervals of 5 and 15 min with R^2 results of 0.232 and 0.384, respectively. These additional results verify that higher resolution for the time interval of analysis leads to models with lower quality.

The simple quadratic model has a better quality than the simple linear model for the hourly analysis, but not for the daily analysis. This can be explained because at the hourly analysis the effect of the performance of the HVAC, which is based on a quadratic trend, has a major influence on the total energy consumption.

The multiple linear model with the outdoor temperature and solar radiation offers a better quality for models on both the hourly and daily analysis. When comparing the quality parameters between the simple and multiple linear regression models, it can be noted that the coefficient of determination (R^2 and R^2_{adj}) improves, but the scattering around the model indicated by $RMSE$ slightly deteriorates. This illustrates the importance of using both parameters to analyze the quality of models.

For comparison purposes, an uncalibrated model of the TxAIRE House #1 was developed using the engineering software BEopt [59] that uses EnergyPlus [60] as simulation engine. Information for the geometry was obtained from blueprints and other known parameters were obtained from a visit to the house. All unknown parameters required by the software were left as default values as given by the software for a new standard construction. The weather file used for the simulation was a modified file of the USA_TX_Tyler-Pounds.Field.722448_TMY3.epw available from the EnergyPlus website [60]. The outdoor temperature and global horizontal radiation recorded onsite were used in the modified weather file. Since the simulations are performed in time steps of 10 min and the internal algorithm within EnergyPlus is used to handle the weather variables during the computation, the outdoor temperature and global horizontal radiation data points do not match with the data in the modified weather file. Therefore, the same data of the two variables from the software simulations are also used in the statistical models to obtain results for model comparison. Table 4 shows the quality parameters obtained for the statistical models and the BEopt software. It can be noted that the coefficient of determination and the RMSE for the hourly statistical models are better than the ones for the engineering model. However, all models are fairly similar for the daily data. Since statistical models avoid the burden associated with the collection of information needed to develop engineering models, the comparison illustrates that the statistical models can be a cost-effective approach to forecast energy consumption in residential buildings in a reasonable and accurate manner.

7. Conclusions

Regression analysis is one of the statistical methods used for developing models for prediction of energy consumption in buildings.

This paper presents relevant information to understand and apply linear regression analysis for application on the residential sector with focus on whole-building energy consumption in single-family homes. The energy signatures and conditional demand analysis are also discussed for better understanding of the use of practical applications of regression analysis for residential energy consumption prediction. The literature review of papers dealing specifically with residential energy consumption using regression analysis supports the feasibility of this statistical approach for model development. The basics of simple and multiple linear regression analysis were applied to data from the TxAIRE Research and Demonstration House #1, as a case study, to illustrate an example of results from regression analysis. As illustrated from the results of the case study, as the time interval of the observed data increases, the quality of the models improves. This is explained by the fact that for longer time periods, the discrepancies among individual effects in shorter time periods are averaged over longer time periods. The solar radiation as a second predictor variable shows improvement of the coefficient of determination, but deteriorates the root mean square error, which justifies the importance of using both parameters to assess the quality of the model based on the developer's criteria. Since HVAC systems accounts for a large portion of the total energy consumption of buildings, and because the performance of HVAC systems can be modeled as a second order polynomial, a quadratic regression model can offer better results for shorter time intervals such as an hour. This is not necessarily true for longer time periods such as a day because the quadratic trend of the HVAC system is lost.

Appendix A

See Table A1 here.

Table A1
Case study's daily energy consumption and weather parameters.

June	Total Electricity (Wh)	Tdb (°C)	GHR (Wh/m2)
1	17,528	27.6	5836
2	13,268	23.2	6303
3	14,813	23.3	8318
4	18,033	25.7	7771
5	25,155	26.6	7230
6	14,006	22.6	1614
7	15,936	22.3	8056
8	14,519	23.1	7429
9	17,298	24.6	5649
10	23,733	27.1	8368
11	20,168	27.3	7801
12	25,913	27.7	7863
13	22,384	28.8	8203
14	29,063	29.8	7510
15	21,933	28.4	5247
16	21,965	28.6	7631
17	23,518	26.5	5359
18	17,174	25.5	2947
19	21,876	25.9	6167
20	21,391	27.9	8043
21	28,139	28.1	7947
22	24,109	27.9	7861
23	25,351	28.2	7579
24	29,439	28.4	7236
25	24,099	28.6	7241
26	30,038	29.2	7860
27	26,026	29.4	8021
28	35,165	31.1	7477
29	23,555	30.1	7883
30	17,988	26.9	8009

References

[1] International Energy Outlook. The U.S. Energy Information Administration. From: ⟨http://www.eia.gov/forecasts/ieo/ieo_tables.cfm⟩; 2013 (retrieved 01.06.14).
[2] Swan Lukas G, Ismet Ugursal V. Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renew Sustain Energy Rev 2009;13:1819–35.
[3] Building America Program, Office of Efficiency and Renewable Energy, the U.S. Department of Energy. Available at: ⟨http://energy.gov/eere/buildings/building-america-bringing-building-innovations-market⟩.
[4] Li Gang, Hwang Yunho, Radermacher Reinhard. Review of cold storage materials for air conditioning application. Int J Refrigeration 2012;35(8):2053–77.
[5] Li Gang, Hwang Yunho, Radermacher Reinhard. Experimental investigation on energy and exergy performance of adsorption cold storage for space cooling application. Int J Refrigeration 2014;44:23–35.
[6] Fumo Nelson. A review on the basics of building energy estimation. Renew Sustain Energy Rev 2014;31:53–60.
[7] Grandjean Arnaud, Adnot Jerome, Binet Guillaume. A review and an analysis of the residential electric load curve models. Renew Sustain Energy Rev 2012;16:6539–65.
[8] Kialashaki Arash, Reisel John R. Modeling of the energy demand of the residential sector in the United States using regression models and artificial neural networks. Appl Energy 2013;108:271–80.
[9] Krüger Eduardo, Givoni Baruch. Predicting thermal performance in occupied dwellings. Energy Build 2004;36:301–7.
[10] Zhu Jiayin, Chen Bin. Simplified analysis methods for thermal responsive performance of passive solar house in cold area of China. Energy Build 2013;67:445–52.
[11] Olofsson Thomas, Sjögren Jan-Ulric, Andersson Staffan.Energy performance of buildings evaluated with multivariate analysis. In: Proceedings of the ninth international IBPSA conference. Montréal, Canada; August 15–18, 2005.
[12] Catalina Tiberiu, Iordache Vlad, Caracaleanu Bogdan. Multiple regression model for fast prediction of the heating energy demand. Energy Build 2013;57:302–12.
[13] Rencher Alvin C, Christensen William F. Methods of multivariate analysis. New Jersey: John Wiley and Sons; 2012.
[14] Chatterjee Samprit, Hadi Ali S. Regression analysis by example. 5th ed. Hoboken, New Jersey: Wiley; 2012.
[15] Weisberg Sanford. Applied linear regression. 4th ed. Hoboken, New Jersey: Wiley; 2013.
[16] Ritz Christian, Streibig Jens Carl. Nonlinear regression with R. New York: Springer; 2008.
[17] Izenman Alan Julian. Modern multivariate statistical techniques: regression, classification, and manifold learning. New York: Springer; 2008.
[18] SPSS software, IBM. Available at: ⟨http://www-01.ibm.com/software/analytics/spss/⟩ (last accessed November 2014).
[19] SAS®, SAS Institute Inc. Available at: ⟨https://www.sas.com⟩ (last accessed November 2014).
[20] SIMCA, Umetrics Inc. Available at: ⟨https://www.umetrics.com/products/simca/⟩ (last accessed November 2014).
[21] STATISTICA, StatSoft Inc. Available at: ⟨https://www.statsoft.com/⟩ (last accessed November 2014).
[22] R, Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien. Available at: ⟨http://www.r-project.org/⟩ (last accessed November 2014).
[23] STATGRAPHICS®, StatPoint Technologies, Inc. Available at: ⟨http://info.statgraphics.com/statgraphics-home⟩ (last accessed November 2014).
[24] NCSS Statistical Analysis and Graphics software, NCSS, LLC. Available at: ⟨http://www.ncss.com/⟩ (last accessed November 2014).
[25] Dormann Carsten F, Elith Jane, Bacher Sven, Buchmann Carsten, Carl Gudrun, Carré Gabriel, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 2013;36(1):27–46.
[26] Tso Geoffrey KF, Guan Jingjing. A multilevel regression approach to understand effects of environment indicators and household features on residential energy consumption. Energy 2014;66:722–31.
[27] Home Energy Yardstick, Energy Star program. The U.S. Environmental Protection Agency. ⟨https://www.energystar.gov/index.cfm?fuseaction=HOME_ENERGY_YARDSTICK.showGetStarted&s=mega⟩ (last accessed May 2014).
[28] Residential Energy Consumption Survey (RECS) End-Use Models FAQs. The U.S. Energy Information Administration. From: ⟨http://www.eia.gov/consumption/residential/methodology/2009/pdf/faqs-enduse-models022013.pdf⟩; 2013 (retrieved 01.06.14).
[29] Hammarsten Stig. A critical appraisal of energy-signature models. Appl Energy 1987;26:97–110.
[30] Rabl Ari, Rialhe Anne. Energy signature models for commercial buildings: test with measured data and interpretation. Energy Build 1992;19:143–54.
[31] Belussi Lorenzo, Danza Ludovico. Method for the prediction of malfunctions of buildings through real energy consumption analysis: holistic and multidisciplinary approach of energy signature. Energy Build 2012;55:715–20.
[32] Yu Frances W, Chan Kam T. Energy signatures for assessing the energy performance of chillers. Energy Build 2005;37:739–46.
[33] Westergren Karl-Erik, Högberg Hans, Norlén Urban. Monitoring energy consumption in single-family houses. Energy Build 1999;29:247–57.

[34] Weather Warehouse. Available at: ⟨http://weather-warehouse.com/⟩ (last accessed May 2014).
[35] Weather data depot. Available at: ⟨http://www.weatherdatadepot.com/⟩ (last accessed May 2014).
[36] Bartels Robert, Fiebig Denzil G. Metering and modeling residential end-use electricity load curves. J. Forecast 1996;15:415–26.
[37] Aigner Dennis J, Sorooshian Cyrus, Kerwin Pamela. Conditional demand analysis for estimating residential end-use profiles. Energy J 1984;5:81–97.
[38] Parti Michael, Parti Cynthia. The total and appliance specific conditional demand for electricity in the household sector. Bell J Econ 1980;11:309–24.
[39] Lafrance Gaétan, Perron Dominique. Evolution of residential electricity demand by end-use in Quebec 1979–1989: a conditional demand analysis. Energy Stud Rev 1994;6(2):164–73.
[40] Aydinalp-Koksal Merih, Ismet Ugursal V. Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector. Appl Energy 2008;85:271–296.
[41] Newshamn Guy R, Donnelly Cara L. A model of residential energy end-use in Canada: using conditional demand analysis to suggest policy options for community energy planners. Energy Policy 2013;59:133–42.
[42] Raffio Gregory, Isambert Ovelio, Mertz George, Schreier Charlie, Kissock Kelly. Targeting residential energy assistance. In: Proceedings of the energy sustainability, July 27–30, 2007. Long Beach, California, USA.
[43] Catalina Tiberiu, Virgone Joseph, Blanco Eric. Development and validation of regression models to predict monthly heating demand for residential buildings. Energy Build 2008;40:1825–32.
[44] Soldo Božidar, Potočnik Primož, Šimunović Goran, Šarić Tomislav, Govekar Edvard. Improving the residential natural gas consumption forecasting models by using solar radiation. Energy Build 2014;69:498–506.
[45] Elsawaf Nehad, Abdel-Salam Tarek, Pagliari Leslie. Evaluation of heat pumps usage and energy savings in residential buildings. Int J Energy Environ 2012;3:399–408.
[46] Min Jihoon, Hausfather Zeke, Lin Qi Feng. A high-resolution statistical model of residential energy end use characteristics for the United States. J Ind Ecol 2010;14:791–807.
[47] Chen Jun, Wang Xiaohong, Steemers Koen. A statistical analysis of a residential energy consumption survey studyin Hangzhou, China. Energy Build 2013;66:193–202.
[48] Schleich Joachim, Klobasa Marian, Gölz Sebastian, Brunner Marc. Effects of feedback on residential electricity demand – findings from a field trial in Austria. Energy Policy 2013;61:1097–106.
[49] Gans Will, Alberini Anna, Longo Alberto. Smart meter devices and the effect of feedback on residential electricity consumption: evidence from a natural experiment in Northern Ireland. Energy Econ 2013;36:729–43.
[50] Mastrucci Alessio, Baume Olivier, Stazi Francesca, Leopold Ulrich. Estimating energy savings for the residential building stock of an entire city: a GIS-based statistical downscaling approach applied to Rotterdam. Energy Build 2014;75:358–67.
[51] Nie Hongguang, Kemp René. Index decomposition analysis of residential energy consumption in China: 2002–2010. Appl Energy 2014;121:10–9.
[52] Bianco Vincenzo, Manca Oronzio, Nardini Sergio. Linear regression models to forecast electricity consumption in Italy. Energy Sources Part B 2013;8:86–93.
[53] Ndiaye Demba, Gabriel Kamiel. Principal component analysis of the electricity consumption in residential dwellings. Energy Build 2011;43:446–53.
[54] Filippína Celina, Ricard Florencia, Flores Larsen S. Evaluation of heating energy consumption patterns in the residential building sector using stepwise selection and multivariate analysis. Energy Build 2013;66:571–81.
[55] Green Button Program. Available at: ⟨http://www.greenbuttondata.org/⟩ (last accessed November 2014).
[56] BEopt™ (Building Energy Optimization). National Renewable Energy Laboratory. Available at: ⟨https://beopt.nrel.gov/⟩ (last accessed November 2014).
[57] Autotune, Oak Ridge National Laboratory. Available at: ⟨http://rsc.ornl.gov/autotune/?q=content/autotune⟩ (last accessed November 2014).
[58] TxAIRE Research and Demonstration Houses. The University of Texas at Tyler. Available at: ⟨http://www.uttyler.edu/txaire/houses/⟩ (last accessed May 2014).
[59] Building Energy Optimization software (BEopt™). National Renewable Energy Laboratory. Available at: ⟨https://beopt.nrel.gov/⟩ (last accessed January 2015).
[60] EnergyPlus, Energy Efficiency and Renewable Energy Office of the U.A. Department of Energy. Available at: ⟨http://apps1.eere.energy.gov/buildings/energyplus/⟩ (last accessed January 2015).