Statistical GIS-based analysis of energy consumption for residential buildings in Turin (IT)

Statistical GIS-based analysis of energy consumption for residential buildings in Turin (IT)

Guglielmina Mutani
Politecnico di Torino
Department of Energy - R3C
Turin, Italy
guglielmina.mutani@polito.it

Roberto Fontana
Politecnico di Torino
Department of Mathematical Sciences
Turin, Italy
roberto.fontana@polito.it

Alison Barreto
Politecnico di Torino
Department of Energy
Turin, Italy
alisonbarreto.fau.ucv@gmail.com

***Abstract*—Greenhouse gas emission is an important issue and the largest source of it is from human activities and from building sectors. Therefore, the building stocks play a key role in the reduction of GHG emissions through the analysis of the energy performance of buildings, in order to understand their behavior and to identify effective models that will allow expanding investigations in vast areas as districts or cities.**

This work analyses space heating energy performance of buildings with a multi-scale approach using the main energy-related variables at building, block of buildings and district scale. The purpose of this study is to identify a simple regression model in order to evaluate the space heating energy consumption of a large part of residential buildings in Turin (IT). A cluster analysis was applied in order to find groups of buildings with similar energy consumptions and to identify the main energy-related characteristics of each group. The analysis was developed with the support of a GIS tool to evaluate the buildings characteristics and a statistical software to identify a stable model at urban scale. The identified models evidenced that the space heating energy consumption not only depends on the characteristics of the building itself, but also on the urban characteristics. At urban scale, the most influential variables were: the heating degree days, positively correlated with the space heating consumption, and the albedo that was negatively correlated. Also, socio-economic variables were utilized: the percentage of working people with a positive correlation and the percentage of young inhabitants with a negative correlation. The statistical GIS-based methodology proposed in this study is simple and then replicable to other urban contexts. This kind of analysis can be useful for policy makers in defining specific energy efficiency measures for each group of buildings to identify new more effective energy performance variables and benchmarks for the different groups of buildings and then to improve the energy performance of a city reducing energy consumptions and the relative GHG emissions.

Keywords—space heating model, residential buildings, linear regression, statistical model, urban scale

I. INTRODUCTION

Even if more than half of the global population now live in cities, the area occupied by them in 2010 only represent 0.5% of the world's surface area and, incredibly, the consumption of this occupied area is 75% of the world's energy consumption [1]. It is estimated that the 68% of the world population will live in cities by 2050, in Italy this percentage will exceed 81% [2]. This assumption with the dawn of environmentalism and concerns regarding resource depletion, the oil crises from 1970s and global climate change brought a discussion on energy consumption especially in high-density urban environments. Consequently, an increasing attention for energy performance of buildings has been given in recent years. As a reaction to climate change, nowadays improving the energy performances of cities has become an important topic in the agenda of governments and decisions makers. In order to build sustainable cities, considerations in urban planning have to be made. The analysis presented in this work had the objective of evaluating the buildings heating energy consumption with a multi-scale approach through the evaluation of the main energy-related variables at different scale. The purpose of this study is to identify a replicable methodology based on multiple linear regression model in order to evaluate the heating energy consumption of buildings with variables at building, block of buildings and urban scale. The analysis was developed using a GIS tool and a statistical software. The GIS tool was used for the association of different databases and the buildings with their energy consumption data, while the statistical software allowed the implementation of different statistical techniques as principal components analysis, multiple linear regression, and cluster analysis in order to evaluate the main energy-related variables and the energy performance models for buildings.

II. LITERATURE REVIEW

Nowadays improving the energy performances of cities has become an important topic in the agenda of governments and decisions makers. During the last years, studies have been implemented trying to understand the main features that influence the energy consumptions at different scales. In 2009, Olofsson et al. [3] analysed the effect of building-specific parameter on energy consumption using ANOVA analysis and PLS-simulation. The PLS model resulted with a better accuracy and it pointed out that the important variables of the buildings were the geometrical characteristics and the construction period. In 2012, Howard et al. [4] studied a model to estimate the building energy end-use intensity for New York city using a robust multiple linear regression. The model was applied to 9 different types of building and it pointed out that the end-use energy depends on building function and not on construction type or building age. In 2014, Mastrucci et al. characterized the building stock of Rotterdam with seven types of dwellings and a bottom-up statistical model was applied to estimate the energy consumption [5]. The model used a multiple linear regression analysis and it pointed out that for electricity consumption number of occupants, floor surface and type of dwelling are the significant variables, while natural gas consumption depends on the floor surface and the type of dwellings. Also in Rotterdam, a study applied a multiple linear regression model on an engineering method to calculate natural gas and electricity consumptions [6]. For this study were considered the average floor area of dwellings, the average number of occupants and the share of dwellings as the main influencing parameters. The energy consumption of only one neighborhood was calculated with an engineering method, while the multiple linear regression model was applied to the entire building stock of the city. The energy consumption derived from the models was compared with real energy consumption, showing a smaller total deviation for the statistical model (5%) and higher for the engineering model (25%). In New York, a study to estimate the building energy-

use intensity was performed by integrating GIS and big data technology at urban scale [7]. Different feature selection strategies and commonly used regression algorithms were included for comparison. Filter, wrapper and embedded methods were performed for the feature selection section and elastic Net, Artificial Neural Network and Support Vector Regression were used for the model. The study concluded that the model built by the Support Vector Regression algorithm on the features selected by Elastic Net had the least cross-validation mean squared error. Normally, researches on buildings energy consumption at urban scale focus on understanding the effect of a single feature on the energy consumption, evidencing the lack of studies that incorporate all the possible features. Therefore, the aim of this study is to identify the most influential variables on building energy consumption at building and urban scale together.

III. THE CASE STUDY OF TURIN (IT)

The case study analysed in this work is the city of Turin, located in the North-West part of Italy. The city has temperate climate, influenced also by the surrounding Alps, with cold-dry winters, warm-humid summers and low wind velocities. Seven weather stations (WS) record the climate variables in different areas of Turin from the city centre to the periphery. Previous studies [8-12] pointed out the main energy-related variables according to weather conditions recorded by the meteorological stations in Turin and its surroundings from the building to the territorial scale. According to the Municipal Technical Map (2015), Turin is characterized by a dense and compact city centre composed by building with similar heights, while the periphery can be represented by more buildings sprawl with irregular urban form and various heights. The major part of the building stock is residential with a mean height of about 14 m, a surface to volume ratio (S/V) between 0.35 to 0.73 and predominant periods of construction 1918-45 and 1961-70 [8-10]. Turin is characterized mainly by residential buildings with central heating systems supplied by the district heating network and natural gas, with an occupation rate higher than 85% and with buildings in a good maintenance condition [13, 14]. In this work, the annual space heating consumptions for space heating of 1,621 residential buildings in Turin for 2/3 heating seasons were analysed in order to understand which are the main influencing variables at building and block of buildings scale. The microclimate variations in the buildings surrounding was analyzed considering the same building typologies located in different areas with different urban morphology, solar exposition and with various outdoor surface materials. Considering buildings with complete information, 1,278 residential buildings were selected with energy consumption data for different heating seasons (from 2009-10 to 2015-16) and for at least 2/3 consecutive heating seasons. The energy consumptions were georeferenced with a GIS tool using data from the Municipal Technical Map of Torino; WSs; socio-economics, urban and buildings characteristics. The energy consumption data of buildings were also normalized on a typical heating season according to the heating degree days registered by the nearest WS. The energy consumption data of residential buildings were normalized. This normalization was on the heating season, which was closer to the average HDD at 20°C of the last 10 years (the 2011-12 heating season was chosen).

In Tables 1 and 2 the analyzed residential buildings in Turin were classified in homogeneous groups considering the period of construction, the surface to volume ratio (S/V) and the number of buildings (the buildings of Turin are mainly old and compact and therefore for the more recent periods of construction it was not possible to complete the classes of S/V and EP_h; see the cells "-"). The greatest part of the analyzed residential buildings was built before the first law on energy savings for buildings L. 373/76. In particular, the 33% of buildings was built in 1961-70, the 25% in 1918-45, the 23% in 1946-60 and the 15% in 1971-80; only the 4% was built after 1981 with some energy efficiency measures. Moreover, many buildings have a low value of surface to volume ratio, so there are principally compact condominiums. In Table 2, the energy consumption for space heating EP_H is reported for Turin; the average values of EP_H increase with the S/V and increase up to 1971-80 and decrease as in accordance with literature [10, 14].

TABLE I. SURFACE TO VOLUME RATIO (S/V) ANALYSIS FOR THE HOMOGENEOUS GROUPS OF BUILDINGS.

Period of construction	Classes of S/V [m^2/m^3]															
	A				B				C				D			
	avg	max	min	n.	avg	max	min	n.	avg	max	min	n.	avg	max	min	n.
< 1945	0.30	0.32	0.25	64	0.36	0.41	0.32	214	0.45	0.59	0.41	108	0.98	1.95	0.59	23
1946-60	0.34	0.38	0.25	227	0.42	0.50	0.38	128	0.81	1.28	0.59	11	-	-	-	-
1961-70	0.29	0.32	0.24	141	0.35	0.38	0.32	217	0.42	0.50	0.38	167	0.80	1.07	0.51	28
1971-80	0.32	0.37	0.24	140	0.41	0.51	0.37	91	0.81	1.19	0.61	7	-	-	-	-
1981-90	0.36	0.50	0.28	41	-	-	-	-	-	-	-	-	-	-	-	-
1991-01	0.40	1.05	0.29	12	-	-	-	-	-	-	-	-	-	-	-	-
>2001	0.40	0.44	0.36	2	-	-	-	-	-	-	-	-	-	-	-	-
buildings	**627**				**659**				**312**				**51**			

TABLE II. SURFACE TO VOLUME RATIO (S/V) ANALYSIS FOR THE HOMOGENEOUS GROUPS OF BUILDINGS.

Period	Classes of EP_h [$kWh/m^3/y$]															
	A				B				C				D			
	avg	max	min	n.	avg	max	min	n.	avg	max	min	n.	avg	max	min	n.
< 45	36.6	86.9	16.4	64	38.9	77.2	17.3	214	41.9	80.3	19.6	108	53.2	65.0	20.4	23
46-60	39.3	98.9	19.9	227	39.4	89.4	17.6	128	46.3	98.9	27.7	11	-	-	-	-
61-70	40.3	82.7	24.0	141	42.0	91.6	19.7	217	42.4	94.1	19.3	167	36.0	68.9	23.5	28
71-80	46.5	99.0	2.2	140	47.9	99.8	26.1	91	52.4	95.3	35.5	7	-	-	-	-
81-90	47.1	89.3	27.8	41	-	-	-	-	-	-	-	-	-	-	-	-
91-01	41.3	58.8	29.2	12	-	-	-	-	-	-	-	-	-	-	-	-
>01	30.5	31.6	29.4	2	-	-	-	-	-	-	-	-	-	-	-	-
build.	**627**				**659**				**312**				**51**			

IV. MATERIAL AND METHODS

A GIS-based methodology to characterize the energy performance (EP) of Turin's buildings heritage has been developed with a bottom-up approach. The accuracy of the models depends on the reliability of the data. For a big city as Turin, the large amount of data missing at urban scale, as the level of renovation of buildings and the renewable energy technologies connected, could cause errors and, in some studies, a correction coefficient could be used to improve the precision of the results [8, 10]. The energy consumption of buildings can be analyzed through different statistical techniques and procedures. The software used was SAS Enterprise Guide version 7.1 with a database composed by 2,230 observations and 80 variables. The following steps summarize the methodological framework of this study:

1. Describe the statistical distribution of energy consumption data.
2. Specify the geometrical, thermos-physical characteristics and systems efficiencies of the buildings and determine and associate the urban characteristics of the surroundings to the buildings using a GIS tool.
3. Use univariate and multivariate analysis techniques (principal component analysis) to improve data quality.
4. Identify a multiple linear regression models.

A. Statistical distribution of energy consumption data

Considering the typical heating season 2011-12, space heating data distributions were analysed considering 17 homogeneous groups of buildings described in Tables 1 and 2. For every homogeneous group, a statistical analysis was performed in order to evaluate the frequency distribution of energy consumption data; in particular, the Normal, Log-Normal and Gamma distributions have been evaluated. Two statistical tests were used in conjunction with the distributions to observe the trend of energy consumptions data and identify the anomalous data: the Kolmogorov-Smirnov (KS) and the chi-squared (χ^2) tests.

B. Buildings and urban variables

With a GIS tool the energy consumption of more than 1,600 residential buildings were georeferenced to combine the energy-use data with the characteristics of the buildings and their surrounding with a statistical approach. All the available data about buildings and urban context were collected and associated to each building. A study on the homogeneous groups of buildings was performed considering the type of distribution of energy performance values for each group.

C. Univariate and multivariate analysis techniques

In the first part of the study univariate techniques were applied. In the following part, the results of the use of multivariate techniques for describing data and understanding the relationship among them are presented. Statistical distributions were used to analyse the database about energy consumptions of different type of buildings and to remove anomalous data; the Pearson coefficient was used to evaluate the grade of correlation of each variable with the energy consumption data.

D. Multiple regression models

This statistical technique is the most used and simple for investigating and modeling the relationship between a dependent and two or more independent variables. The heating energy consumption was estimated using a multiple linear regression, which is expressed by:

$$Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_p \cdot x_{ip} + \varepsilon_i \quad (1)$$

where Y_i is dependent variable, the space heating energy consumption, x_{ij} are the independent variables, β_j are the parameters estimated and ε_i is the random error of each observation i, i=1,…,N. The standard assumptions for the errors ε_i are that they are independent and normally distributed with mean 0 and constant variance σ^2, $\varepsilon_i \sim IIND(0, \sigma^2)$.

The model in Eq (1) was identified using Ordinary Least Squares (OLS) method. It follows that the observed values yi (i=1,…,N) can be written as:

$$y_i = b_0 + b_1 \cdot x_{i1} + \cdots + b_p \cdot x_{ip} + e_i \quad (2)$$

where b_j are the least squares estimates of β_j (j=0,1,…,p) and e_i (i=1,…,N) are the residuals. The predicted values $\hat{y}_i$ are computed as $b_0+b_1 \cdot x_{i1}+\cdots+b_p \cdot x_{ip}$ (i=1,…, N).

A meaningful result from a multiple regression model can be obtained by confirming at least approximately the assumptions of the method. The validity of the assumptions is checked using some diagnostic tools on the residuals. Residuals should be approximately normal, with constant variance (homoscedasticity) and uncorrelated. The distribution of the residuals was evaluated through probability graphs. The "residuals vs predicted values" and the "residuals vs regressors" graphs have been also built. The homoscedasticity of the residuals was also evaluated by the White test that considers the two following hypotheses:

$$H_0: \sigma_i^2 = \sigma^2 \quad (3)$$

$$H_1: \exists_{i,j}\ such\ that\ \sigma_i^2 \neq \sigma_j^2 \quad (4)$$

where the null hypothesis (Eq. 3) represent the equal variance (σ^2) for the errors while the alternative hypothesis (Eq. 4) the different variance for the errors. In the case of a non-constant variance for the errors a variance-stabilizing transformation is required in order to obtain more accurate parameter estimators of the model.

A potential problem for the validity of the results of a multiple regression is the collinearity among regressors. Variance Inflation Factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. In a multiple regression model, multicollinearity makes difficult to test how much the independent variables affects the dependent variable since they are all influencing each other. The multicollinearity was assessed by the VIFs calculated by the reciprocal of the inverse of R_j^2 (R^2 is the coefficient of determination) of an independent variable x_j as it is expressed by:

$$VIF_j = \frac{1}{1-R_j^2};\ j = 1, \ldots, p. \quad (5)$$

Usually VIF values greater than 5 suggest that the regression coefficients are poorly estimated. To evaluate the adequacy of the model, the coefficient of determination R-squared (R^2) and the adjusted R^2 (Adj R^2) were used. R^2 represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Similarly, the adjusted R^2 is a modified version of the coefficient of determination that has been adjusted for the number of predictors in the model. While the value R^2 increases with the addition of variables in the model, the adjusted R^2 increases only if the new term improves the model more than would be expected by chance. Standard regression output includes the results of the statistical tests which compare the null hypothesis H_0 against the alternative hypothesis H_1 where H_0 and H_1 are defined as:

$$H_0: \beta_j=0;\ H_1: \beta_j \neq 0. \quad (6)$$

Usually the null hypothesis is rejected (in other words the parameter is significant) when the T statistic, or t-ratio (defined as the ratio between b_j, the estimate of β_j, and the estimate of the standard deviation of the estimator B_j of β_j), is large or more precisely when the relative p-value is less than 5%. In this work, the estimate of the standard deviation of the estimator B_j of β_j is simply denotes by s_{B_j}; it follows then t-ratio is equal to: b_j/s_{B_j}.

For this study, 80 variables were considered and the stepwise method was performed in order to select only the most influential variables. It is an automatic selection procedure which combines forward selection and backward elimination methods. The backward elimination step considers a subset of regressors and for each regressor the t-ratio is computed. If the smallest absolute value of the t-ratios is less than a prespecified value, the corresponding regressor is eliminated. The forward selection step is similar but considers adding instead of eliminating variables. A new variable is added if the corresponding t-ratio is the largest and its value is greater than a prespecified value. To avoid the effect of the unit of measures the independent variables have been standardized. Considering the mean $\bar{x}_j$ and the standard deviation σ_j of the value $\{x_{ij}: i=1,\ldots, N\}$, the standardization is defined as:

$$z_{ij} = \frac{x_{ij}-\bar{x}_j}{\sigma_j}. \quad (7)$$

V. RESULTS AND DISCUSSION

The steps described in the previous paragraph about the methodological framework were used also to describe the results of this study.

A. Statistical distribution of energy consumption data

In literature was found that energy performance (EP) of residential buildings depends mainly by their period of construction and surface to volume ratio (S/V). Then a study on the energy performance distribution was completed to identify the characteristics of the homogeneous group of buildings with similar energy performances. In this study, it was found that space heating consumption data distributions were not normal but Gamma and Log-Normal, as can be seen from Table 3. The trend Log-Normal was also found in [15] for office buildings because energy consumption densities are always non-negative, which indicates that they may be Log-Normally distributed. The frequency distribution is very closed to the theoretical or expected distribution. In Table 3 the statistical analysis values with 3 ranges of data is reported and in only 3 cases the KS test led to reject the hypothesis of Gamma/Log-N distributions (in red).

TABLE III. ANALYSIS OF STATISTICAL DISTRIBUTIONS GAMMA AND LOG-NORMAL FOR THE HOMOGENEOUS GROUPS OF BUILDINGS

Period of construction	S/V classes	S/V_{avg} m^2/m^3	$EP_{h,avg}$ $kWh/m^3/y$	*Statistical test*			
				Gamma distr.		*Log-N distr.*	
				X^2 (p-values)	KS (deviation)	X^2 (p-values)	KS (deviation)
< 1945	A	0.30	36.64	19.4%	0.17	35.4%	0.17
	B	0.36	38.93	8.9%	0.09	6.6%	0.09
	C	0.45	41.91	12.1%	0.13	12.0%	0.13
	D	0.98	53.16	8.7%	0.09	29.0%	0.28
1946-60	A	0.34	39.29	12.5%	0.09	38.4%	0.09
	B	0.42	39.38	37.1%	0.12	21.9%	0.12
	C	0.81	46.32	71.2%	0.41	15.5%	0.41
1961-1970	A	0.29	40.35	6.4%	0.11	34.1%	0.11
	B	0.35	41.98	19.8%	0.09	27.7%	0.09
	C	0.42	42.40	26.9%	0.11	17.4%	0.11
	D	0.80	36.03	20.8%	0.26	45.1%	0.26
1971-80	A	0.32	46.47	8.0%	0.12	14.0%	0.12
	B	0.41	47.87	32.5%	0.14	13.7%	0.14
	C	0.81	52.41	85.5%	0.51	11.1%	0.51
1981-90	A	0.36	47.11	46.0%	0.21	13.1%	0.21
1991-2001	A	0.40	41.33	44.6%	0.39	38.7%	0.39

B. Characteristics and systems efficiencies of the buildings and urban characteristics of the surroundings

Before applying any statistical procedures is necessary, first analyzed the data for understanding the tendency, the dispersion, the accuracy also to detect outliers. The sample used in this study counts with 2,230 observations that represent residential buildings located in different part of the city. The variables present in the dataset can be classified into three main groups:

1. Building variables as: energy consumption, type of use, period of construction, maintenance level, area and volume, type of envelope (with thermal transmittances of opaque and transparent envelope), orientation, compactness or S/V and type of heating system.
2. Urban variables representing the surroundings where each building is located as (at block of buildings scale): climate and microclimate variations, buildings density, buildings coverage ratio, buildings average height, percentage of heated volume, aspect ratio, distance from the center of the city, main buildings and streets orientation, solar exposition and albedo coefficient of urban environment (Figure 1).
3. Socio-economic variables: number of inhabitants, family total components, percentage of males and females, percentage of foreigners, age, educational level, employment rate, dwelling property, workforce people and income level.

Buildings characteristics were defined at building scale, while the urban variables and socio-economic factor at block of buildings scale. Then, these last variables were defined with average values at block of buildings scale and then associated at every building in that area. From the analysis of this dataset, mainly considering the average and the standard deviation values of the variables, the residential database counted on 2,060 observations after removing the anomalous observations (i.e. not residential buildings with high volumes or energy consumption and unoccupied buildings with very low energy consumption).

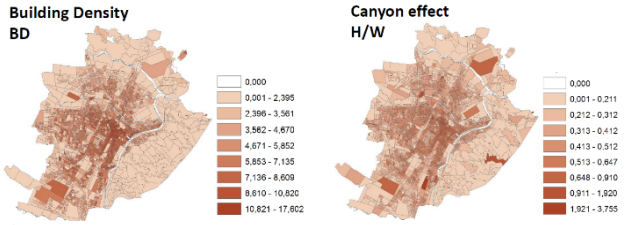

Fig. 1. Analysis of urban variables in the city of Turin: BD and H/W ratio

C. Univariate and multivariate analysis techniques

Through the univariate analysis, it is possible to describe the observations by looking only one variable at the time. The high correlation between energy consumption and heated volume is represented. The coefficient of Pearson r = 0.898, close to the value of 1 especially for low volumes, measures the linear dependency between these two variables. Multivariate analysis was applied after, considering more than one variable at the time. This type of analysis was made through principal components and correlation analyses performed using the statistical software SAS. The aim of these analyses was to understand which variables are more correlated with energy consumption and which are not independent of each other, a necessary condition for the linear regression model to be identified in the following paragraph.

D. Perform a multiple linear regression models

Considering the model represented by equations (1), (2) and (7) the following multiple regression models were performed. The first stepwise method (model '1') resulted in a model with 24 independent variables exposing that the most influential variables are at building scale the volume, the heated volume, the net area and the transmittance of the walls. Table 4 display the general information of the models showing that the selected variables are capable of explaining the 84% of the variance in the energy consumption. In addition, Table 6 shows the information obtained from the model and its respective values for the estimates b_j of the parameters β_j, the estimates of standard errors s_{B_j}, t-ratios and the variance inflation factor (VIF). It can be observed that the VIF values for the volume and net area are very high because they are correlated to the heated volume; then they will not be considered.

The distribution of the residuals are shown by a histogram (Figure 2) and a probability graph (Figure 3). The residuals do not accomplish the normality assumption of a multiple linear regression model, presenting a Kernel distribution with long tails especially for high values of the response. Also, looking at the probability graph it can be seen that the residuals cannot be considered as normally distributed. The scatterplot of the residuals displayed in Figure 4 exposed a funnel pattern

meaning that the variance of the residuals tend to increase with an increasing of the predicted value so the variance of the residuals is not constant going against to homoscedasticity assumption. Table 5 displays the results from the White test, evidencing that the residuals are not homoscedastic and are then heteroskedastic (p-value is smaller than 0.05). As the assumptions of a multivariate linear regression analysis were rejected, a transformation on the dependent variable was performed using "log10" function. Therefore, the stepwise method using the "log10" of the dependent variable was performed resulting in a model with 26 variables, an adjusted R^2 value of 0.816 (model '2'). The residuals of this model showed a better distribution even if the p-value of the White test is very close to 5%. For this model, the scatterplot "residuals versus predicted" evidences the presence of a quadratic trend. The best result was found to be the "log10" including the quadratic terms z_{ij}^2 corresponding to each predictor z_{ij} (models '3' and '4'). For this reason, the square of each predictor was added to the model and then the multiple linear regression resulted in a polynomial expression that is summarized in the following equation:

$$Log_{10}(Y_i) = \beta_0 + \beta_1{}^{(l)} \cdot z_{i1} + \beta_1{}^{(sq)} \cdot z_{i1}{}^2 + \cdots + \beta_p{}^{(l)} \cdot z_{ip} + \beta_p{}^{(sq)} \cdot z_{ip}{}^2 + \varepsilon_i \qquad (8)$$

TABLE IV. MODEL INFORMATION WITH EQUATION (1)

Model	R^2	Adj R^2
'1'	0.84	0.84
'2'	0.818	0.816
'3'	0.877	0.874
'4'	0.88	0.87

TABLE V. WHITE TEST FOR HOMOSKEDASTICITY

Model	DF	χ^2	Pr > ChiSq
'1'	324	337.22	0.029
'3'	877	850.26	0.7354

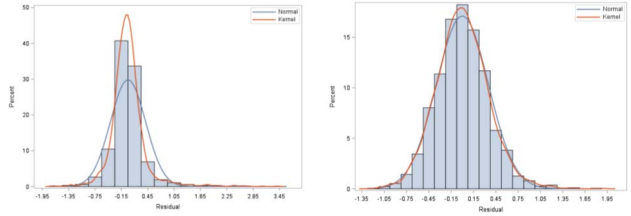

Fig. 2. '1' and '3' model: Normal and Kernel distribution of errors

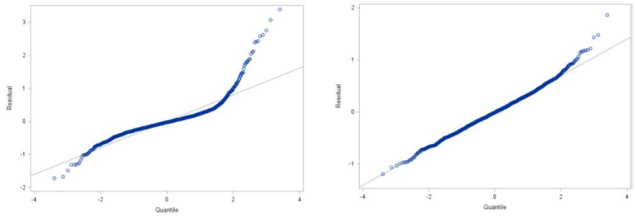

Fig. 3. '1' and '3' model: probability graph

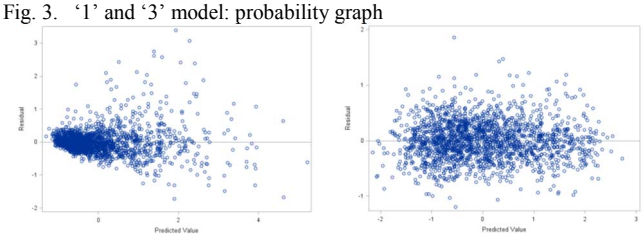

Fig. 4. '1' and '3' model: variance distribution of the residuals

The stepwise method (model '3') resulted in an improvement in comparison to the previous models. The model achieves an adjusted R-square of 0.874 with 41 variables, as it is displayed in Table 4. The biggest influence on EP (energy performance index) is given by the heated volume, the area and the number of floors (Table 7). While these variables implied an increase of the energy consumption the net area, the squared residents and the squared area present strong negative influence on the dependent variables. The VIF in Table 4 evidence that some variables are still correlated with other variables inside the model. In Figures 2 and 3, are displayed the graph related to the residuals showing a better result with the transformation although there are still some difficulties for predicting higher values. The results for evaluating the heteroscedasticity of the residuals can be analysed through Table 7 and Figure 4; the scatterplot does not evidence any pattern on the residuals while the White test does not reject the null hypothesis of having homoscedastic residuals.

TABLE VI. MODEL 1 WITH EQUATION (1)

Variable (z_{ij})	b_j (parameter estimate of β_j)	Standard Error S_{B_j}	t-ratio	p-value Pr > |t-ratio|	VIF
Intercept	2.53E-10	0.009	0	1	-
Volume	-0.612	0.09	-6.82	<0.0001	92.63
University degree	-0.082	0.017	-4.89	<0.0001	3.22
Independent system	-0.052	0.011	-4.71	<0.0001	1.4
Floor	-0.045	0.016	-2.9	0.004	2.78
BOsc	-0.034	0.01	-3.37	0.001	1.16
Rented dwellings	-0.027	0.012	-2.25	0.025	1.6
Distant center	-0.025	0.016	-1.53	0.126	2.97
Age 5-9	-0.023	0.01	-2.23	0.026	1.23
BHsc	-0.023	0.014	-1.64	0.102	2.2
Age 25-29	-0.021	0.012	-1.82	0.069	1.53
R Women	-0.02	0.011	-1.85	0.064	1.28
Condition Status Poor	-0.017	0.01	-1.61	0.108	1.25
Solid fuel	0.021	0.01	2.18	0.029	1.09
Buildings used	0.024	0.01	2.42	0.016	1.08
Age 65-69	0.024	0.011	2.22	0.026	1.34
Electric Energy	0.025	0.01	2.59	0.01	1.06
Year 1	0.037	0.011	3.37	0.001	1.38
Age more than 74	0.043	0.014	3.14	0.002	2.12
Workforce people (WP)	0.046	0.017	2.75	0.006	3.14
HDD	0.062	0.011	5.45	<0.0001	1.5
Window HTS	0.083	0.026	3.22	0.001	7.63
Wall HTS	0.282	0.036	7.74	<0.0001	15.26
NET area	0.335	0.081	4.15	<0.0001	74.74
Heated volume	0.862	0.022	39.9	<0.0001	5.37

A more stable model was found but it still has some problems that need to be solved as the presence of collinearity and influential observation that may change the model. The presence of multicollinearity inside a model may dramatically affect its usefulness; it implies a linear dependence among the variables making the regression coefficient poorly estimated. The solution to this problem was to remove the variables that may have collinearity with others. The removal of the variables cannot be made at the same time but one by one because the VIF of the remaining variables could change. The model '3' showed that some of the independent variables have linear dependency. The variable with the biggest value of VIF is the square of the family total components; then, the complete model was performed with the remaining variables. The removal of the variable "Family components2" resulted in a decrease on the VIF value for the square of the resident variables. The procedure was repeated 3 times in total until was reached a model without highly linear dependence among the independent variables. The variables that were removed from the model were: the square of the gross volume, the net area and the squared of the family total components. Finally, the correlated independent variables and the most influential observations were removed and all the assumptions were achieved. The last model '4' counts of 32 predictors which are both linear and quadratic; VIF values are always less than 5. Tables 4 and 7 present the information about the last model. From the model '4', in Figure 5, emerged that the heated volume, the area and the S/V have the highest influence on the energy consumption; as expected, these variables are directly proportional with energy consumption. Also urban parameters

and the socio-economic variables influence the energy consumptions.

TABLE VII. MODEL 4 WITH EQUATION (8)

Variable (z_{ij})	b_j (parameter estimate of β_j)	Standard Error S_{B_j}	t-ratio	p-value Pr > |t-ratio|
Intercept	0,19	0,02	7,78	<,0001
Age 25-29	-0,04	0,01	-4,08	<,0001
Age 45-49	-0,02	0,01	-2,46	0,014
Age 5-9	-0,05	0,01	-5,01	<,0001
Age 60-64	-0,04	0,01	-4,25	<,0001
Age10-14 2	-0,01	0,00	-2,36	0,0184
Age65-69 2	0,01	0,00	2,11	0,0351
Albedo	-0,04	0,01	-2,99	0,0028
Albedo 2	0,02	0,01	2,72	0,0066
Area	0,2	0,02	9,42	<,0001
Area 2	-0,07	0,01	-8,53	<,0001
BHsc	-0,03	0,01	-2,3	0,0213
Bosc 2	-0,01	0,00	-2,77	0,0057
Families in rented dwellings	-0,07	0,01	-4,62	<,0001
Floor	0,14	0,02	8,33	<,0001
HDD 2	0,05	0,01	6,75	<,0001
Heated volume	1,01	0,03	38,46	<,0001
Heated volume 2	-0,14	0,01	-15,87	<,0001
Height 2	-0,05	0,01	-5,46	<,0001
HWsc 2	0,03	0,01	4,11	<,0001
Natural gas 2	-0,03	0,01	-4,36	<,0001
Other fuels	0,03	0,01	2,63	0,0085
Residential buildings 2	-0,02	0,01	-2,4	0,0165
Residents 2	0,02	0,01	2,58	0,0098
S/V	0,18	0,02	10,67	<,0001
S/V 2	-0,02	0,00	-5,17	<,0001
UM	-0,03	0,01	-2,3	0,0217
University degree 2	0,02	0,01	2,08	0,038
University degree	-0,07	0,02	-4,35	<,0001
Workforce P 2	-0,03	0,01	-4,89	<,0001
Workforce people	0,04	0,01	2,89	0,0039
WP Occupied 2	0,03	0,01	6,2	<,0001
Year 1	0,04	0,01	3,79	0,0002

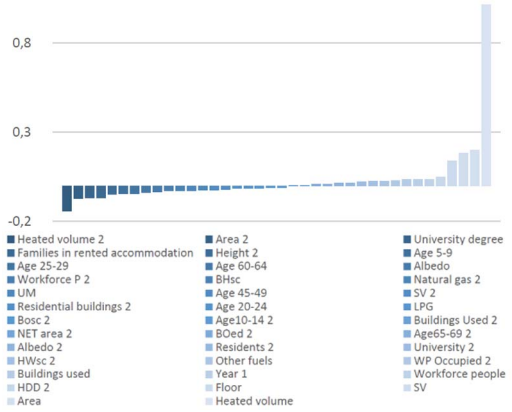

Fig. 5. Weight of the variables on the EP of residential buildings.

VI. CONCLUSIONS

The multiple linear regression model was applied for residential buildings and was validated using a group of observations selected randomly through the software SAS. The best results were found with the "log10" of the dependent variable (energy consumption) and including the quadratic terms of each independent variable with coefficients of determination of 0.82-0.88; the model is good, also for less amount of utilized variables, after all the verifications (model '4'). The model for residential building points out that at building scale the most influential variables were the heated volume, the area and the S/V ratio (higher values for these variables are associated with a higher energy consumption). Regarding to the urban parameters the model evidences that the HDD, the aspect ratio H/W and the albedo are positively correlated. Contrary, the buildings coverage ratio and the building height are negatively correlated with the total energy consumption. The number of residents inside a census section is associated with an increase on energy consumption but these variables does not showed the stronger influence. About the socio-economic variables, the university degree of the inhabitants was the variable that showed a stronger influence on energy consumptions and its relationship was actually negative. Also, the percentage of families living in rented dwellings are associated with a decrease on energy consumption, while the work-force and the occupied work-force are associated with an increase of the energy consumption. For the variables that characterize the energy vector, only four variables were important: the natural gas, the LPG, the percentage of residential buildings and the percentage of buildings used. The greater decrease on energy consumption is obtained when the natural gas is more used as it is associated with new boilers or district heating heat exchangers. In conclusion, the resulted models evidenced the most important variables that affect the space heating consumption of buildings evidencing that energy-use does not only depends on the characteristics of the building itself but also on the urban characteristics of the surrounding environment. The statistical methodology used in this study could be used for evaluating the energy consumption at urban scale in another city considering also the characteristics of the population besides the physical aspects as the features of buildings and of the urban context. All of these aspects are important for urban planners, architects and decisions makers for identifying the best solution when retrofit interventions or an energy plan are needed, considering the real effects that the decisions will have on the energy consumption and GHG emissions for a district, city or territory. The use of other statistical methods, like probabilistic graphical models, could be investigated in future researches [16].

REFERENCES

[1] Schneider A., Friedl M.A., Potere D. (2009). A new map of global urban extent from MODIS data, Environm. Research Letters,10.1088/1748-9326/4/4/044003.

[2] United Nations, DESA/Population Division 2018. World Population Prospects.

[3] Olofsson T., Andersson S. and Sjögren J.U. (2009). Building energy parameter investigations based on multivariate analysis, Energy and Buildings 41(1), 2009.

[4] Howard B., Parshall L., Thompson J., Hammer S., Dickinson J., Modi V. (2012). Spatial distribution of urban building energy consumption by end use, Energy and Buildings 45, 2012.

[5] Mastrucci A., Baume O., Stazi F. and Leopold U. (2014). Estimating energy savings for the residential building stock of an entire city: A GIS-based statistical downscaling approach applied to Rotterdam, Energy and Buildings 75, 2014.

[6] Nouvel R., Mastrucci A., Leopold U., Baume O., Coors V.and Eicker U. (2015). Combining GIS-based statistical and engineering urban heat consumption models: Towards a new framework for multi-scale policy support, En. and Buildings 107.

[7] Ma J. and Cheng J. C. (2016). Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology, Applied Energy 183.

[8] Delmastro C., Mutani G., Pastorelli M., Vicentini G. (2015). Urban morphology and energy consumption in Italian residential buildings, IEEE Conference Publications, DOI: 10.1109/ICRERA.2015.7418677.

[9] Mutani G., Fiermonte F. (2016). Microclimate Models for a Sustainable and Liveable Urban Planning, Urban and Landscape Perspectives 19, 10.1007/978-3-319-51535-9.

[10] Mutani G., Todeschi V. (2017). Space heating models at urban scale for buildings in the city of Turin (Italy), Energy Procedia, 122, 2017, DOI: 10.1016/j.egypro.2017.07.445.

[11] Carozza M., Mutani G., Coccolo S., Kaempf J. H. (2017). Introducing a hybrid energy-use model at the urban scale: the case study of Turin (IT), 3rd BSA-Italy Conference, pp 209-216, ISSN: 25316702.

[12] Torabi Moghadam S., Toniolo J., Mutani G., Lombardi P. (2018). A GIS-Statistical Approach for Assessing Built Environment Energy Use at Urban Scale, Sustainable Cities and Society 37, DOI: 10.1016/j.scs.2017.10.002.

[13] Guelpa E., Mutani G., Todeschi V., Verda V. (2017). A feasibility study on the potential expansion of the district heating network of Turin, Energy Procedia, Vol. 122, CISBAT 2017, DOI: 10.1016/j.egypro.2017.07.446.

[14] Mutani, G., Pastorelli, De Bosio, F., A model for the evaluation of thermal and electric energy consumptions in residential buildings: The case study in Torino (Italy), ICRERA 2015, 10.1109/ICRERA.2015.7418637.

[15] Li X., Yao R., Li Q., Ding Y., Li B. (2018). An object-oriented energy benchmark for the evaluation of the office building stock, Utilities Policy 51, 1-11.

[16] Coscia C., Fontana R., Semeraro P. (2018). Graphical models for complex networks: an application to Italian museums, Journal of Applied Statistics 45.