Clustering Analysis of Electricity Consumption of Municipalities in the Province of Pichincha-Ecuador Using the K-Means Algorithm

Oscar F. Núñez-Barrionuevo(✉), Edilberto A. Llanes-Cedeño, Javier Martinez-Gomez, Jorge I. Guachimboza-Davalos, and Jesús Lopez-Villada

Universidad Internacional SEK, Quito, Ecuador
ofnunez.mee@uisek.edu.ec

Abstract. This work shows an analysis of electricity consumption by parishes in the province of Pichincha to define consumption affinities between them. To do this, categorical variables are encoded to be used with the unsupervised machine learning algorithm k-means. As relevant results, four types of consumption are established and they are analyzed in relation to population density while algorithm represents 100% of the variability of the data, and municipalities with different consumption trends are identified.

AQ1

Keywords: Power consumption · Clustering · K-means

1 Introducción

Global warming has become one of the biggest concerns of governments around the world, so efforts are increasingly being made to counteract its effects and mitigate its impact on the population. In virtue of this, numerous studies have been carried out where it has been shown that the polluting gases generated by industries, transport and many of the habits of the population have been the cause of this phenomenon. Adverse events such as fires, floods and droughts are only part of the effects of climate change on the planet [1]. One of the main causes is emissions of polluting gases, including CO_2. These emissions must be controlled immediately before the consequences are irreversible for the population, so both the issuance of public policies and the search for new renewable energy sources must be proposed by governments for their mandatory application in the areas of the industry [2]. Furthermore, these effects of global warming are related to the modification of the electricity consumption of the population suffering extreme temperatures, causing an increase in the use of heating systems [3]. Therefore, one of the ways of acting to limit and prevent its environmental, social and economic consequences, related to the increase in temperature, is to reduce energy

M. Botto-Tobar et al. (Eds.): ICCIS 2020, AISC 1273, pp. 1–9, 2021.
https://doi.org/10.1007/978-3-030-59194-6_16

consumption. Since the current generation, transport and consumption model is absolutely dependent on fossil fuels. In this way it is unsustainable as a consequence of the climate change that it implies.

In Latin America, there is an average of 0.65 tons of CO_2 for each MWh generated [4,5]. In Ecuador, this result is 397.5 g of CO_2 per KWh, where the highest value corresponds to 2010 since the generation of energy from non-renewable sources was (52.2%) [6]. However, with the change in the productive matrix, 60.85% of electricity production is based on renewable energy where 58.53% is related to Hydroelectric sources [7]. This production is oriented 30.93% to residential consumption and 26.01% to the Industrial sector [6]. However, the generation of electricity through power plants causes changes in ecosystems, the disappearance of species, soil erosion, among others. For this reason, it is necessary to carry out a process of recognition of consumption parameters that allow discovering people's habits. This process is carried out in order to have information for all market participants to cooperate effectively in saving electricity consumption. In addition, it provides information to power generation plants to carry out proper planning and avoid overloading their equipment [8].

With the constant increase in household electricity demand, the analysis of electrical energy consumption (EEC) is increasingly important. Therefore, the acquisition and analysis of data has become a widely used tool these days since it allows to determine the behavior of EEC in a precise way. In this sense, the use of user tags (identifiers) and the appropriate technology can provide a more intuitive and concise expression to EEC analysis compared to traditional analysis [9]. For this reason, the use of machine learning algorithms allows us to recognize more relevant patterns of the phenomenon by studying based on complex mathematical calculations that seek to emulate certain functionality of the human brain, such as grouping into sets of objects in relation to their most characteristic evident. In this sense, by having information on electricity consumption by cantons and parishes, the grouping criterion allows them to be organized in such a way that subgroups with attributes similar to each other, but different from others, are formed. This process is framed within unsupervised learning since there is only one set of input data that we must obtain information about the structure of the output domain, which is not available [10]. With this, an EEC can be established by sectors of the country and defined electricity consumption policies. However, the data acquisition process can be a challenging task. This is because most government databases have mostly categorical information. Consequently, it significantly limits data analysis [11]. For this, there are techniques that allow encoding variables to convert them to a numerical system. However, it is necessary to be careful in the use of these attributes since they can oversize the machine learning algorithm and provide erroneous information when generating the model. It must be emphasized that the geopolitical distribution is divided by provinces, cantons, and municipality [6].

For this reason, the present work exposes the development of clusters within the energy consumption of Ecuador in the province of Pichincha to group them into population areas with similar consumption trends related to their canton

and municipality. For this, a data preprocessing stage is performed to encode the categorical variables and eliminate those that do not present relevant information to the machine learning algorithms. Subsequently, technical criteria are used to determine the number of groups to be carried out and finally establish electricity consumption trends. As a result, this study presents an acceptance of 100% that explains the variability of each group.

The rest of the document is structured as follows: Sect. 2 presents the related works. Section 3 indicates the methodological scheme of data analysis. Section 4 presents the cluster analysis to provide the grouping of electricity consumption. Section 5 shows the results obtained with the different error analyses. Finally, section6 presents the results and future work.

2 Related Works

Works such as [8–10] have presented important contributions related to the grouping in the behavior of electricity consumption of residential customers, in order to understand the personalized demands of the user and provide them with specific services. However, consumption patterns can be significantly different in relation to geographic location, population density, consumption habits, among others. In addition, most of the works use a supervised analysis criterion where the output variables are known. In this case, the aim is to find non-existent grouping parameters at a glance. That is why there are pending problems, mainly oriented in the Ecuadorian population since there is a lack of these studies and above all, they do not use machine learning algorithms.

3 Materials and Methods

This section shows, on the one hand, data acquisition (Sect. 3.1), normalization (Sect. 3.2), the selected grouping algorithm (Sect. 2.3) and the data analysis scheme (Sect. 3.3). AQ2

3.1 Data Acquisition

The data collection is through the information presented by the Ministry of Electricity and Renewable Energy in its monthly reports during the every year. In these reports, the company that distributes electricity to the provinces, cantons, and municipalities is explained in detail. In this specific case, the consumption information for the years 2019 and 2018 is taken. All the data is stored in the array $\mathbf{Y} \in \mathbb{R}^{m \times n}$ where m represents the number of samples and n the attributes of the database. In this case, $m = 2{,}550$ and $n = 11$. Attributes refer to: month, canton, municipality, type of equipment with a voltage at 220, number of customers, energy billed, increase in consumption, residential consumption, subsidy energy generated, billing of the service and value of the subsidy.

3.2 Database Normalization

Matrix normalization means adjusting the measured values on different scales with respect to one in common. This process is carried out prior to making mathematical models [7]. With this, it is avoided that there are variables that, according to their nature of the data, contribute in a large weighted amount to the model as they have very high values in relation to the rest. In addition, it prevents the use of categorical variables from being diminished due to its low scale (0 to 1). That is why there are different methods and forms of standardization. In this case, the standard normalization given by the following formula is used:

$$\frac{X - \mu}{\sigma} \tag{1}$$

3.3 Clustering

One of the important computational challenges is having the ability to recognize characteristics to group similar elements. The `K-means` method aims to partition a set of m observations into k groups. Where each value of m belongs to a group of k whose mean value of the distance is the closest. To carry out this process, the following process is established:

Algorithm 1. K-means pseudo-code

Input: Dataset
Output: each m assigned to a group k
1: Choose the number of k fo clusters
2: Select at random K points the centroids
3: Assign each data point to the closest centroid (That forms k clusters)
4: Compute and place the new centroid of each cluster
5: Reassign each data point to the new closest centroid
6: **if** any reassignment took place **then**
7: go to step 4
8: **else**
9: go to Fin
10: **return:** Each m on k cluster

However, the randomness of the values of k can cause different forms of grouping. Therefore, it is necessary to properly define its value in order to group only the data that contains the largest number of similar attributes [12]. Consequently, the `K-means ++` version allows us to eliminate this problem by analyzing a set of observations $(x_1, x_2, ..., x_p)$, where each observation is a real vector $d - dimensional$, the grouping of $K - means$ aims to divide the p observations into $k(<= p)$ sets $S = S_1, S_2, ..., S_k$ to minimize the sum of squares within the group (WCSS) (that is, the variance). This can be seen in the next equation:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \tag{2}$$

where ν_i is the point average in S_i.

3.4 Data Analysis Scheme

The present work focuses on grouping the municipalities of the Pichincha province by electricity consumption habits. For this, a data analysis model is required that represents everything previously discussed, this is shown in Fig. 1.

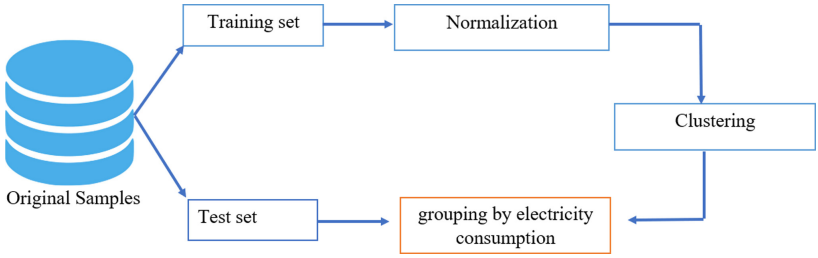

Fig. 1. Data analysis scheme proposed

4 Results

Analysis (WCSS) is performed with a value of $p = 10$ and determine the appropriate value of k. Based on the variance, $k = 4$ is determined to be ideal relative to the attributes of the $\mathbf{Y}$ database. This can be seen in Fig. 2.

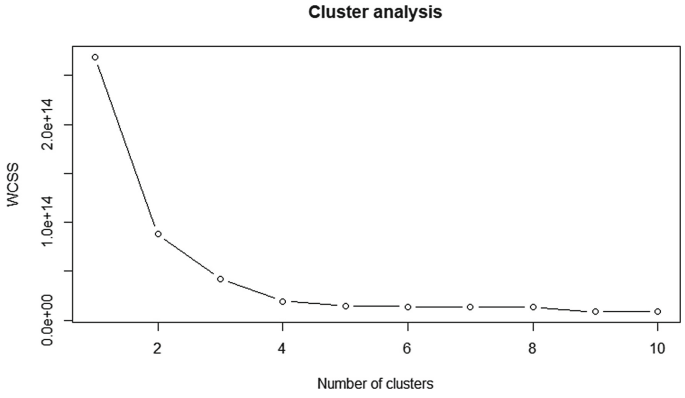

Fig. 2. WCSS analysis to select k value

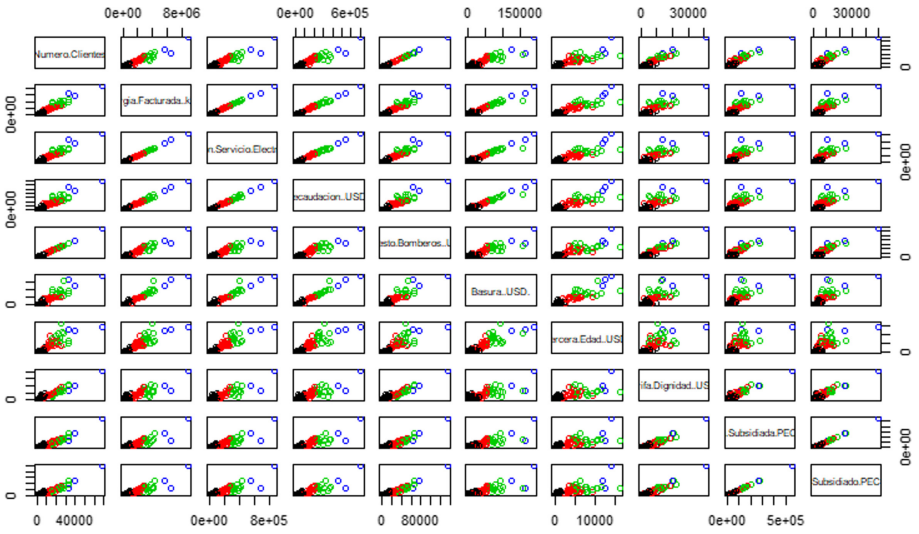

Fig. 3. Data variance analysis to represents the fit the model

In order to know the percentage of acceptance of the model in relation to the distribution of the data. It is done in the analysis of variance. This is together with the implementation of the WCSS model. As a result, there is a normal distribution of the data to the model with an acceptance of 99.9%. The multi-dimensional summary of the analysis of variance can be seen in Fig. 3.

The implementation of the algorithm allows grouping by cantons in relation to consumption in the province of Pichincha-Ecuador. As main information, the centroid values of each cluster are established. Where the electric consumption has the values of: Excessive consumption: 7037116.0 Kwh, high: 3458488.1 Kwh, average: 1984133.6 Kwh and low: 295444.3 Kwh. In order to graphically observe the organization of the clusters, it is carried out by reducing the dimensionality to two dimensions by means of the Principal Component Analysis (PCA) algorithm Fig. 4.

With the machine learning algorithm developed, the municipalities of each canton are grouped in relation to the type of consumption previously established. A summary can be made by cantons of the province of Pichincha. This can be seen in the Table 1.

The Metropolitan District of Quito has the highest electricity consumption in the province. However, its municipalities have great variability in consumption. With the aim of having a geographic analysis, the form of consumption can be established with colors: Excess consumption (red), high (tomato), average (yellow) and low (green). This can be seen in Fig. 5. It can be seen that the parishes of Calderón, Conocoto, and Iñaquito have higher consumption in relation to the rest of the municipalities (around 7037116.0 Kwh). This is due to their high population density and they are much higher than the rest. However,

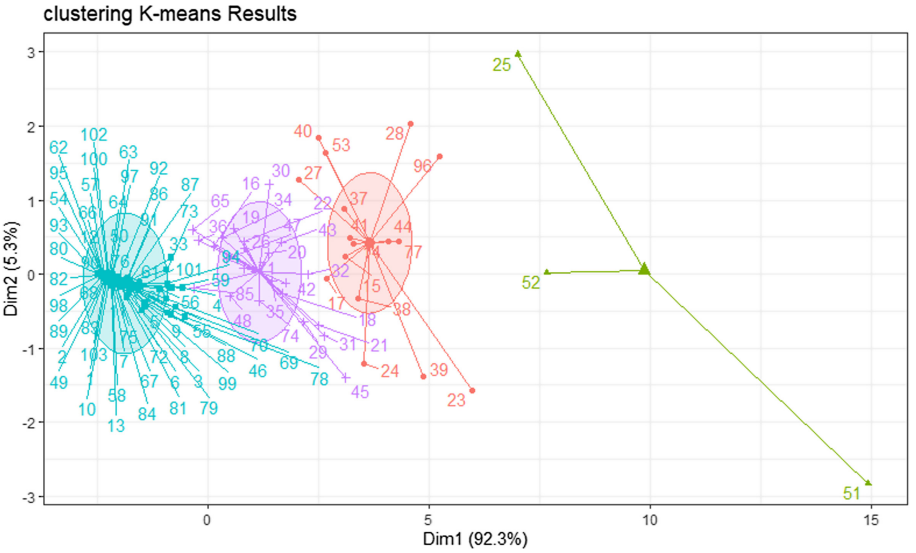

Fig. 4. K-means results for power consumption in Pichincha

Table 1. Summary of grouping of municipalities by type of electricity consumption

Power consumption clustering	Municipalities
EXCESSIVE	IÑAQUITO(25), CALDERON (CARAPUNGO)(51), CONOCOTO(52)
HIGH	BELISARIO QUEVEDO(14), CARCELÉN(15), COCHAPAMBA(17), EL CONDADO(23) GUAMANÍ(24), JIPIJAPA(27), KENNEDY(28), PONCEANO(37), PUENGASÍ(38), QUITUMBE(39), RUMIPAMBA(40), SAN BARTOLO(41), SOLANDA(44), CUMBAYA(53), TUMBACO(77), SANGOLQUI(96)
MODERATE	CENTRO HISTóRICO(16), COMITÉ DEL PUEBLO(18), COTOCOLLAO(19), CHILIBULO(20), CHILLOGALLO(21), CHIMBACALLE(22), ITCHIMBIA(26), LA ARGELIA(29), LA CONCEPCIóN(30), LA ECUATORIANA(31), LA FERROVIARIA(32), LA MAGDALENA(34), LA MENA(35), MARISCAL, SUCRE(36), SAN ISIDRO DEL INCA(42), SAN JUAN(43), TURUBAMBA(45), ALANGASI(47) AMAGUAÑA(48), NAYON(65),POMASQUI(71), SAN ANTONIO(74), MACHACHI(85)
LOW	TANDAPI(1), SAN MIGUEL DE LOS BANCOS(2), PUERTO QUITO(3), CAYAMBE(4) JUAN MONTALVO(5), CANGAHUA(6), OLMEDO (PESILLO)(7) SAN JOSE DE AYORA(8), TABACUNDO(9), LA ESPERANZA(10) MALCHINGUI(11), TOCACHI(12), TUPIGACHI(13), LA LIBERTAD(33), QUITO(46), ATAHUALPA (HABASPAMBA)(49), CALACALI(50), CHAVEZPAMBA(54), CHECA (CHILPA)(55), EL QUINCHE(56), GUALEA(57), GUANGOPOLO(58), GUAYLLABAMBA(59), LA MERCED(60), LLANO CHICO(61), LLOA(62), NANEGAL(63), NANEGALITO(64), NONO(66), PACTO(67), PERUCHO(68), PIFO(69), PINTAG(70), PUELLARO(72) PUEMBO(73), SAN JOSE DE MINAS(75), TABABELA(76), YARUQUI(78), ZAMBIZA(79), CAYAMBE(80), ASCAZUBI(81), CANGAHUA(82), OTON(83), SANTA ROSA DE CUZUBAMBA(84), ALOAG(86), ALOASI(87), CUTUGLAHUA(88), EL CHAUPI(89), SAN RAFAEL(95),(TANDAPI)(90), TAMBILLO(91), UYUMBICHO(92), MALCHINGUI(93), SANGOLQUÍ(94), COTOGCHOA(97), RUMIPAMBA(98), SAN MIGUEL DE LOS BANCOS(99), PEDRO VICENTE MALDONADO(101),PUERTO QUITO(102), MINDO(100)

Iñaquito has an average population density (around 18 thousand people) but with high electricity consumption. On the other hand, parishes such as El Condado or San Antonio have a high population density (25,000 and 30,000 people respectively) do not have excessive electricity consumption. This happens in the same way with the municipalities of El Quinche and Guayllabamba, which has around 17 thousand people and have low electricity consumption (average of 295444.3 Kwh).

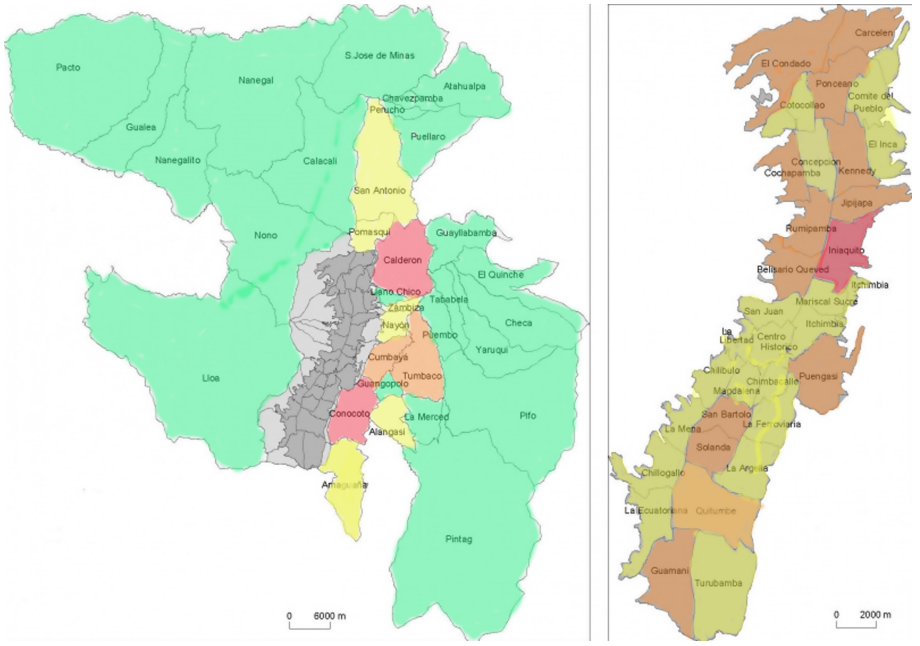

Fig. 5. Cluster distribution of municipalities in Pichincha providence. Excess consumption (red), high (tomato), average (yellow) and low (green).

5 Conclusions and Future Works

This process of grouping by electricity consumption in contrast to the population density can be deduced that parishes with 15 thousand inhabitants have a consumption of more than 17,000 Kwh per month. However, some of them do not present this similar trend and it is necessary to carry out a more in-depth analysis of the characteristics of electricity consumption.

It can be seen in the graph that sectors with similar geographical locations and population density can change the habits of electricity consumption. This can be seen specifically in the high and medium consumption sets. With this, better government policies can be presented for better planning of electricity production.

As future work, it is proposed to make use of a decision-making support tool to have an adequate interface on consumer behavior with monthly and annual reports.

References

1. Huang, X., Wang, S.: Prediction of bottom-hole flow pressure in coalbed gas wells based on GA optimization SVM. In: *Proceedings of 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2018*. Institute of Electrical and Electronics Engineers Inc, pp. 138–141 December 2018
2. He, L., Song, Q., Shen, J.:K-NN numeric prediction using bagging and instance-relevant combination. In: *Proceedings - 2nd International Symposium on Data, Privacy, and E-Commerce, ISDPE 2010*, pp. 3–8 (2010)
3. Bose, S., Goel, A., Shankar, T., Mageshvaran, R., Rajesh, A.: Energy efficient heterogeneous network with daily load variation. In: *2017 Innovations in Power and Advanced Computing Technologies, i-PACT 2017*, vol. 2017-January. Institute of Electrical and Electronics Engineers Inc, pp. 1–4 January 2018
4. Zhang, X.M., Grolinger, K., Capretz, M.A., Seewald, L.: Forecasting residential energy consumption: single household perspective. In: *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*. Institute of Electrical and Electronics Engineers Inc, pp. 110–117 January 2019
5. Diao, L., Sun, Y., Chen, Z., Chen, J.: Modeling energy consumption in residential buildings: a bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. Energ. Build. **147**, 47–66 (2017)
6. Ministerio de Electricidad y Energía Renovable – Ente rector del Sector Eléctrico Ecuatoriano. http://historico.energia.gob.ec/
7. Parra Narváez, R.: Factor de emisión de CO₂ debido a la generación de electricidad en el Ecuador durante el periodo 2001–2014, *Avances en Ciencias e Ingeniería*, vol. 7, no. 2, December 2015
8. Wang, Y., Chen, Z., Xu, Z., Gang, G., Lu, J.: User electricity consumption pattern optimal clustering method for smart gird. In: *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 567–570 August 2018
9. Zhong, C., Shao, J., Zheng, F., Zhang, K., Lv, H., Li, K.: Research on electricity consumption behavior of electric power users based on tag technology and clustering algorithm. In: *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 459–462 July 2018
10. Kangping, L., Wang, F., Zhao, Z., Zengqiang, M., Hongbin, S., Chun, L., Bo, W., Jing, L.: Analysis on residential electricity consumption behavior using improved k-means based on simulated annealing algorithm. In: *2016 IEEE Power and Energy Conference at Illinois (PECI)*, pp. 1–6 February 2016
11. Toapanta-Lema, A., Gallegos, W., Rubio-Aguilar, J., Llanes-Cedeño, E., Carrascal-García, J., García-López, L., Rosero-Montalvo, P.D.: Regression models comparison for efficiency in electricity consumption in ecuadorian schools: a case of study. In: *Applied Technologies*, Botto-Tobar, M., Zambrano Vizuete, M., Torres-Carrión, P., Montes León, S., Pizarro Vásquez, G., Durakovic, B., Eds. Cham: Springer International Publishing, pp. 363–371 (2020)
12. Rezaei, S., Sharghi, A., Motalebi, G.: A framework for analysis affecting behavioral factors of residential buildings' occupant in energy consumption, *J. Sustain. Archi. Urban Des.* **5**(2), 39–58 (2018). http://jsaud.sru.ac.ir/article_895.html%0Ahttp://jsaud.sru.ac.ir/pdf_895_9f6e3441f7399851f2185e37696ed98e.html

Author Queries

Chapter 16

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	
AQ2	Please note that the section 2.3 is cited in the text but not provided. Kindly check and confirm.	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	New matter followed by
Insert in text the matter indicated in the margin		or (x2)
Delete	through single character, rule or underline or through all characters to be deleted	or (x2)
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	