Data Compression and Sampling Period of Daily Energy Consumption

A. Carolina Flores, G. Fraidenraich [*]

***Abstract*—Data compression will be extremely necessary in order to solve the problem of large data volume generated by smart meters. The aim of this work, is to investigate the source coding theory, which establishes the entropy as a fundamental limit on the performance of conventional data compression algorithms applied to daily load curve of a typical resident consumer. We have proposed the applying of Pulse Code Modulation+Huffman and Differential Pulse Code Modulation+Huffman as methods to represent the consumption as few bits are possible, and we have explored the performance of compression when the typical sampling period of 15 minutes is reduced. Furthermore we have determined the second and third order entropies as a limit to data source compression schemes.**

***Index Terms*—Smart Grid, Smart Meter, data compression, entropy.**

I. Introduction

The electrical network has presented almost no advances when energy metering technology is considered. Traditionally, the power grid refers to interconnected transmission system using analog technology. Although humans are well equipped for analog communications, it is not particularly efficient for data storing and data transmission. Today the smart grid is widely considered to be the next-generation of a supposed digital electricity grid. The new platform is devised to use smart meters as a measurement device of electricity flow from the energy utility to the customers, and vice versa. Of course, this new platform has received enormous attention, because it offers many options and capabilities that are not possible with traditional meters. Energy utilities companies, all around the world, are starting to use smart meter technology [1], [2].

The smart meter is a digital meter that is located at home or business to measure the amount of electricity used and stores it over short intervals. It remotely sends this information from different networks to the central operation point, as shown in Fig. 1 [3]. However, since the whole population will shortly start using this new device, a huge amount of data is expected to be transmitted and stored over time. Utilities will have to solve the data collection problem, storage challenges, and learn how to analyze and act based on this new information. Accurate demand forecasting is essential to energy planning and trading.

This is the reason for this work, whose aim is to use data compression techniques on the native format of energy consumption, in order to alleviate the cost of data transmission. The proposed process is sketched in Fig. 3, which will be explained along the paper.

Moreover, the traditional sample period of 15 minutes is analyzed and as will be shown, we prove that this is not the appropriated time in order to avoid the information loss. All the analysis is based on the entropy of the energy demand curve.

Although, the compression scheme for data networks is an old solved problem, to the best of these author's knowledge, this problem has never been investigated to daily load data in the literature.

For a better understanding, the paper is organized as the follow. The Section II describes the scheme proposed and it presents the important concepts to use for understanding of the process. The Section III details the algorithm used to create of daily consumption readings and the mains parameters. The Section IV shows the results obtained and its analysis the performance. Finally, the Section V presents the conclusions reached about the project.

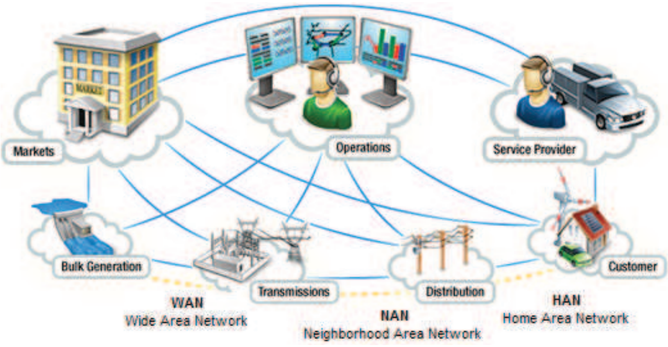

Fig. 1. Smart Grid Infrastructure. Source: [3].

II. Data Compression Model

The analog daily load curve defined as $x(t)$, as show in Fig. 2, it represents the consumption of a typical residence along the day. Normally, the peak D_M of this curve is around the time that most of the domestic devices are used.

In the new smart meter devices, $x(t)$ will be sampled at time kT, $k \in$ integers. In principle, according to the Nyquist criteria, the correct value of T would produce no loss of information. In the sequel, the equivalent discrete sampled load curve x_T will be quantized. This quantization block inserts, depending on the number of used bits, certain distortion. The output of quantization block can be represented either in PCM (pulse code modulation) format or DPCM (differential pulse

[*] A. Carolina Flores, G. Fraidenraich , Department of Communications, School of Electrical and Computer Engineering, University of Campinas, Campinas-SP, Brazil, E-mails: carolina@decom.fee.unicamp.br, gf@decom.fee.unicamp.br. This work was partially supported by CAPES.

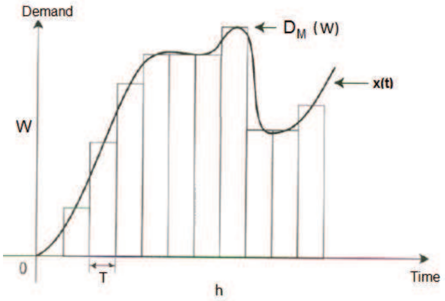

Fig. 2. Daily load curve (KW x h)

code modulation) format. After the quantization block, the signal is then compressed using Huffman algorithm [10]. All the steps are shown in Fig. 3.

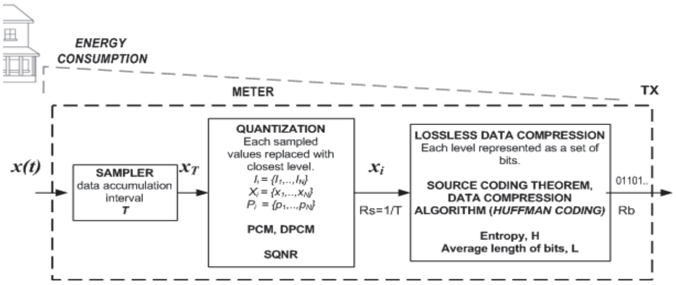

Fig. 3. Block diagram of data compression.

A. Quantization

The digital processing of daily load signals x_T, where the source alphabet is not discrete, requires an infinite number of bits. Therefore x_T must be quantized to its discrete representation x_i with a finite number of levels. PCM and DPCM belong to this class.

The sampled signal is divided into N non overlapping regions called quantization intervals $I_i, i = 1, 2, .., N$ [10], [13] and the number of bits required to represent each codeword would be $\log_2 N$. All quantization intervals I_i are of equal length and with these definitions, the following can be stated:

$$x_T \in I_i \leftrightarrow Q\left(x_T\right) = x_i \tag{1}$$

Since we will be interested on the distribution of the levels x_i, we defined the probability of occurrence of level x_i as p_i, in such a manner that $\sum_{i=1}^{N} p_i = 1$. This quantization process is a lossy data compression, because the decompressed data is not exactly the same as the original data. Instead, some amount of distortion D is tolerated [10], [13]. The distortion is defined in a usual way as:

$$D = E[(x_T - x_i)^2] = \int_{-\infty}^{\infty} (x_T - x_i)^2 f(x_T) dx_T \tag{2}$$

where $f(x_T)$ denotes the probability density function of the source random variable x_T.

The signal-to-quantization-noise ratio (SQNR) is defined as:

$$SQNR|_{dB} = 10 \log_{10} \frac{E[{x_T}^2]}{D} \tag{3}$$

In particular, PCM quantizes the sampled signal x_T directly in x_i. Unlike PCM, the quantizer DPCM is designed to quantize the differences between two consecutive samples. Also DPCM uses a predictor to produce an estimate at each step [13]. The schematic of a PCM and DPCM systems are shown in Fig. 4, and Fig. 5, respectively. DPCM system can be described by the following equations:

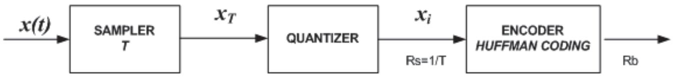

Fig. 4. Schematic of a PCM System.

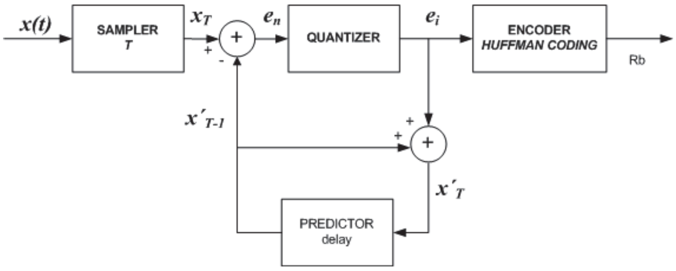

Fig. 5. Schematic of a DPCM System.

$$e_T = x_T - x'_{T-1} \tag{4}$$

$$e_i = Q[e_T] \tag{5}$$

$$x'_T = x'_{T-1} + e_i \tag{6}$$

where, x'_{T-1} is the predicted reconstructed value; x'_T is the reconstructed value; e_T and e_i are the prediction error and its quantized value, respectively [13].

B. Lossless Data Compression

The source coding theory sets as a fundamental limit on the performance of all data compression algorithms [11], the first order entropy H_1 is given by:

$$H_1 = \sum_{i=1}^{N} p_i \log_2 \left(\frac{1}{p_i}\right) \qquad (bits/quantized\ \ value) \tag{7}$$

which is a function of the first-order distribution p_i.

When all the levels are equally likely to occur (i.e., $p_i = 1/N$), the first order entropy is given by:

$$H_1 = \log_2 N \qquad (bits/quantized\ value) \tag{8}$$

The fundamental source coding theorem states that the average length of a code L is defined by:

$$L = \sum_{i=1}^{N} l_i p_i \qquad (bits/quantized\ \ value) \tag{9}$$

where l_i is the individual codeword length and according to entropy theorem $L \geq H_1$ [10].

The entropy defined in (7) is appropriated for random variates. However for random process, the appropriated metric would be the entropy defined as $\mathcal{H}$ [12]. For random process which can be represented by a Markov chain of second-order, the second-order entropy is defined as:

$$\mathcal{H}_2 = \sum_{j=1}^{N} p_j \sum_{k=1}^{N} P_{k|j} \log_2 P_{k|j} \tag{10}$$

where $P_{k|j}$ is the conditional probability that the present sample is x_k given that the previous sample was x_j.

In a similar way, if the random process is better described by a Markov Chain of third order, the third order entropy rate is defined as:

$$\mathcal{H}_3 = \sum_{j=1}^{N} p_j \sum_{k=1}^{N} P_{k|j} \sum_{m=1}^{N} \log_2 P_{m|k,j} \tag{11}$$

where $P_{m|k,j}$ is the conditional probability that the present sample is x_m given that the previous sample was x_k and the one before that was x_j.

For a correlated random process, as the daily demand profile process, these metrics would be of interest since it represents the limit of on the number of required bits to represent each sample x_i.

III. Daily load profile Simulator

We obtain the daily load curve of residential consumer using [5], [7] at a sample rate of $1/900\ Hz$, meaning data capture periods of $15\ minutes$. Modeling a residential load curve considers the typical use of some household appliances, such as shower for heating water, microwave for food preparation, refrigerator for food storage, fan for environmental conditioning, lamps, stereo, TV, computer, DVD, washer, iron, etc. The average power (Watts), average duration (min) are based on [8]. The average peak time (hour) is based on [7]. This is shown in Table I.

TABLE I
Appliances - Consumption and Powers

Appliances	Average Power (W)	Average Peak Time (h)	Average Duration (min)
Refrigerator	90	-	-
Shower	3500	7, 20	8
Microwave	1200	9, 12, 19	10
Iron	1000	10, 17	10
Washer	500	9, 17	30
Computer	180	10, 18	300
Fan	120	12	240
TV	110	11, 21	180
DVD	100	22	60
Lamp	100	20	240
Stereo	45	10, 18	240

Therefore, this random simulation is performed in order to take into account the usage habits of the appliance. As an example, the room light is on during the night and normally turned off when residents go to sleep. However, the refrigerator is constantly connected although switches on and off as a function of internal temperature. For the sake of simplicity, this load is considered with a fixed power value along the day.

Based on Table I, it generates a Gaussian random variable for each appliance with mean set as the time of largest use and arbitrary standard deviation (for example 2 hours). Once the random variates is drawn, the table column average duration and average power is respectively employed in order to assign the duration and demand of each appliance.

The simulation considers average demands in the period of $15\ minutes$. For example, suppose that a shower with power of $3.5\ kW$ is turned on for $8\ minutes$. Then, it means an average power in $15\ minutes$ of $3.5 \cdot 8/15 = 1.86\ kW$.

Consequently, the load curve for each equipment is simulated, and the curves are summed up leading to a daily load curve for one consumer, as illustrated in Fig. 6. This is a possible load profile for only one day of consumption and not a profile of average consumption, and this individual consumption is not necessarily the same every day, because there is randomness. For this reason, our simulator performs the average of several curves as shown in Fig. 7.

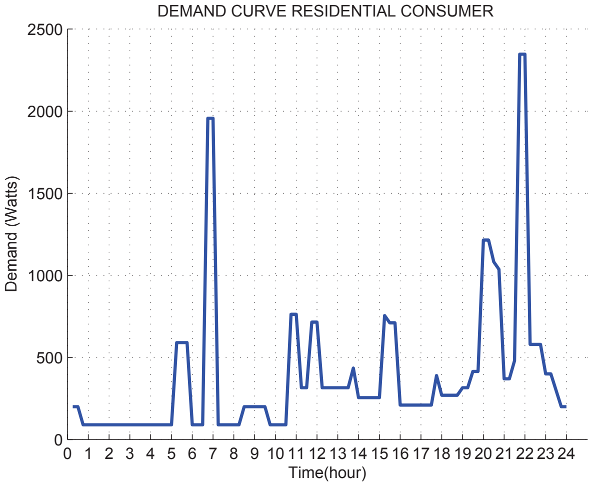

Fig. 6. Daily Load Profile for a single Residential Consumer with a sample time of 15 minutes.

In order to validate our simulator, the resulting simulation curve was compared to real data obtained from [9]. The real measured curve is presented in Fig. 8.

IV. Data Analysis

Now with the simulator and the use of entropy of first, second and third order, it is possible to conclude several important points.

A. Sampling data rate

In the literature, a very well established sampling period of $T = 15$ minutes is largely used. Of course, by increasing the sampling rate we can get any disturbance on the demand curve that could be explored by the network operators. On the other hand, the larger the sampling rate the larger will be the amount of data to be transmitted. So it is clear that there is a trade-off between these two quantities. In order to investigate what

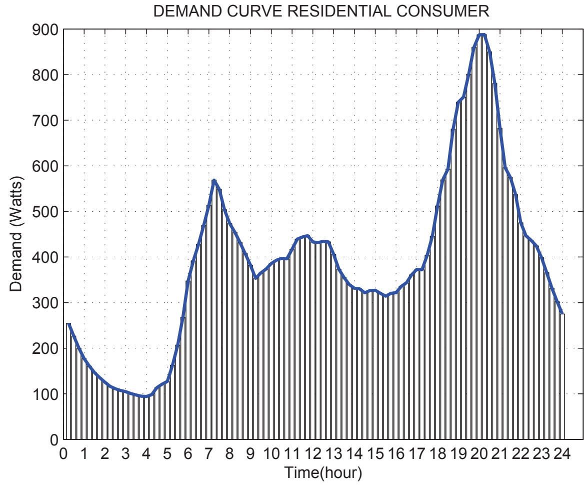

Fig. 7. Average Daily Load Profile for a Residential Consumer with at sampling time T = 15 minutes.

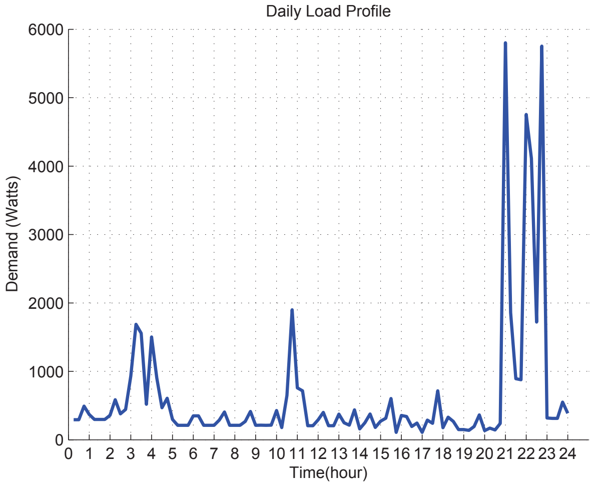

Fig. 8. Real average measured Daily Load Profile for a single Residential Consumer with a sample time of 15 minutes. Source: [9].

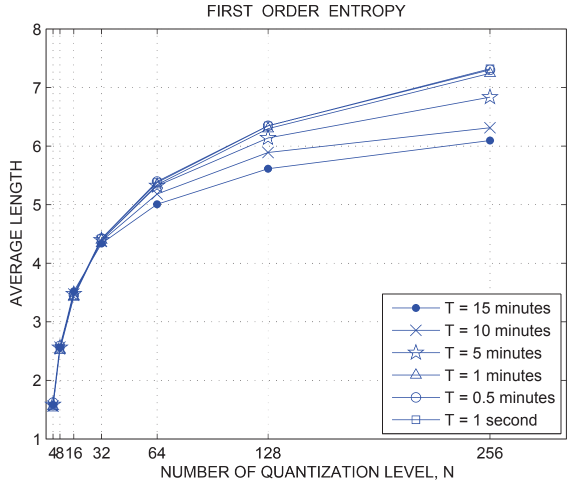

Fig. 9. First order entropy for sampling periods $T = 15$ min, $T = 10$ min, $T = 1$ min, $T = 0.5$ min and $T = 1$ sec.

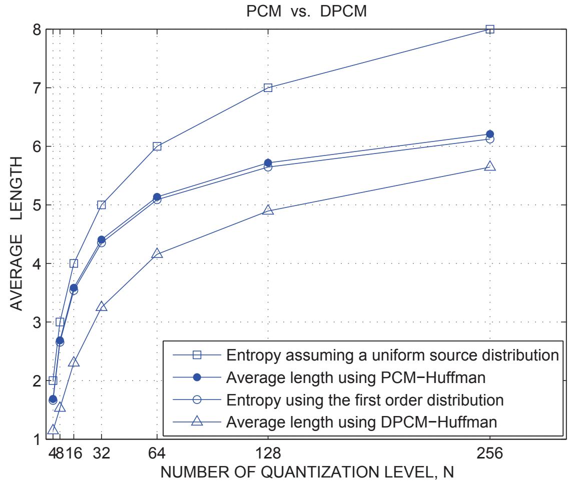

Fig. 10. Entropy using a uniform source distribution and first order distribution, and average number of bit given by Huffman coding, for both PCM and DPCM, at sampling period T = 15 minutes.

would be the ideal sampling rate, Fig. 9 shows how entropy changes with different sampling period. It is noteworthy to point out that there is no significantly change in the entropy for a sampling period less than 1 minute. That is, it is useless to increase the sampling rate above of this number since there will be no gain in entropy. In other words, 15 minutes of sampling period implies in a loss of information.

B. PCM versus DPCM

Using both PCM and DPCM techniques, Fig. 10 shows the average length as a function of the number of quantization levels. As can be seen, in the best case, DPCM reduces about 20% the average length. Note also in this figure, how the PCM-Huffman code approaches the first order entropy, as expected.

This figure is also very useful to estimate the total number of bits and therefore the hard memory to store the information. Since we have T=15 minutes, in one day it will be necessary $24/0.15 = 96$ samples. Considering 256 quantization levels and the use of PCM-Huffman, the total number of bits will be $6 \times 96 = 576$ bits a day. It seems a negligible number, but when we consider a city with 17 millions of people like São Paulo, it would imply in almost 10 billions of bits a day, which is an impressive number! If the same rationale is performed to a DPCM coder, we obtain 9 billions of bits. These number highlight the importance of our analysis.

C. Second and Third Order Entropies for Daily Demand Curves

In order to investigate further the effect of the correlation among the sampled values, we have calculated the second and third order entropies rates as defined in (10) and (11), respectively, as shown in Fig. 11. The results are very impressive since it shows that better compression schemes can be used in order to exploit the correlation. As can be seen, for number of quantization levels equals to 256 and sampling period of

T=1 minute, a code with only two bits per sample would be required, which represents a reduction of 67% when compared to a pure PCM-Huffman scheme.

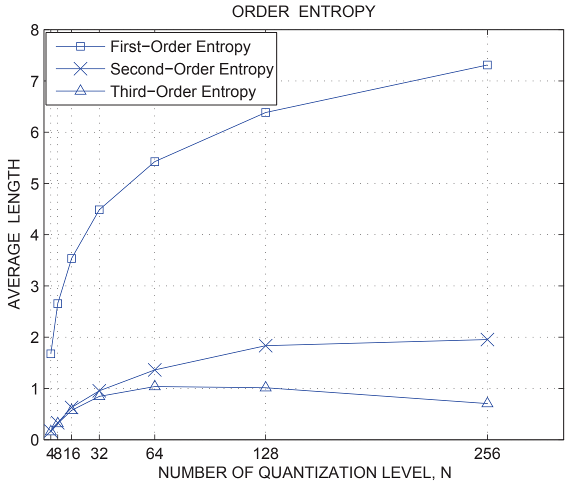

Fig. 11. First-Order Entropy, Second-Order entropy and Third-Order Entropy at sampling period T = 1 minute.

V. Conclusions

We have concluded that DPCM+Huffman was able to reduce the information about 20%. Moreover, choosing an appropriate sample rate for smart meters is a very important decision that may affect in loss of information or amount of data to be transmitted. So, decreasing the sampling period of smart meters to 1 minute will increase the amount of data to be transmitted. This conclusion is not based on the traditional Nyquist criteria, but rather on the entropy of the signal source.

In order to investigate further the effect of the correlation among the sampled values, we have calculated the second and third order entropies rates. The results are very impressive since it shows that better compression schemes can be used in order to exploit the correlation. Using a sampling period of T=1 minute, a code with only two bits can be used, which represents a reduction of 67% when compared to a pure PCM-Huffman scheme.

Therefore, the results of this framework can give us a useful tool to anticipate the data load management effects.

Acknowledgments

Thanks to CAPES, for financial support.

References

[1] James Momoh, *SMART GRID fundamentals of Design and Analysis*. Mohamed, Wiley, IEEE PRESS, 2012.

[2] B. A. Harney, ”Smart Metering Technology Promotes Energy Efficiency for a Greener World,” pp. 1−3, 2009.

[3] IEEE Smart Grid, ”IEEE: The expertise to make smart grid a reality,” April 2013, available at: *http://smartgrid.ieee.org*.

[4] SGIC, ”Smart grid information clearinghouse,” April 2013, available at: *http://www.sgiclearinghouse.org/LearnMore*.

[5] R. De Oliveira, M. G. Zago, and D. S. Gastaldello, ”Algoritmos para estimar curvas de cargas a partir de padres de hbitos de consumo,” pp. 775−781.

[6] J. A. Jardini, C. M. V Tahan, M. R. Gouvea, S. U. Ahn, F. M. Figueiredo, and S. Member, ”Daily Load Profiles for Residential, Commercial and Industrial Low Voltage Consumers,” vol. 15, no. 1, pp. 375−380, 2000.

[7] R. I. O. Executivo and A. N. O. Base, ”Avaliao do mercado de eficincia energtica do brasil,” 2005.

[8] Inmetro, ”Imformao ao consumidor,” April 2013 ,available at: *http://www.inmetro.gov.br/consumidor/tabelas.asp*.

[9] J. Kolter and M. Johnson, ”REDD: The Reference Energy Disaggregation Data Set,” in Proceedings of the SustKDD Workshop on Data Mining Applications in Sustainabili, Abril 2013, available at: *http://redd.csail.mit.edu*

[10] John G. Proakis, Masooud Salehi, Gerhard Bauch, *Contemporary Communication Systems Using MATLAB and Simulink*. BookWare Companion Series TM, 4th edition.

[11] Data-Compression, ”Theory of Data Compression,” April 2013, available at *http://www.data-compression.com/index.shtml*.

[12] Thomas Cover, *Elements of Information Theory*. Willey, Third Edition.

[13] Allen Gersho, Roert M. Gray, *Vector quantization and signal compression*. Kluwer.