Jurnal
Teknologi

Full Paper

MODELING RESOURCE ALLOCATION FOR ENERGY EFFICIENCY IN DATA CENTERS ON THE SMART GRID

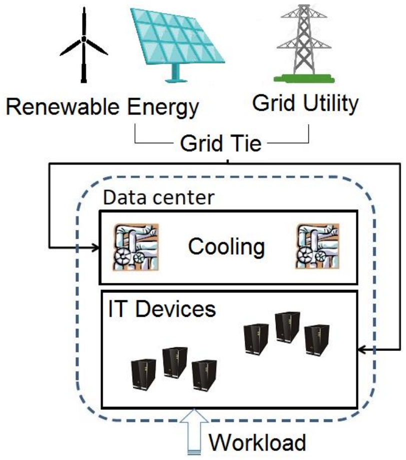
Sergio Mora Martinez, Jhon Edwin Vera*, Jonathan Avendano Perez, Lizet Camila Salgado

Vicerrectoría de Investigación, Universidad ECCI, Cl. 51 # 19-12, Bogotá, Colombia

Article history
Received
17 October 2019
Received in revised form
24 February 2020
Accepted
26 February 2020
Published online
23 April 2020

*Corresponding author
smoram@ecci.edu.co

Graphical abstract



Abstract

Data centers are constantly growing and evolving in number and complexity due to increasing demand for Internet-based services. As a result, energy consumption in data centers has increased significantly in recent years, which has become a critical concern for IT enterprises and governments because of high operational costs and negative environmental impact. Therefore, green solutions are needed to integrate the use of renewable energy with the development of reduction strategies in energy consumption. In this study, we investigated the performance of a system that can simulate a data center constrained by service-level agreements, energy consumption, power generation, and non-exponentially distributed service times. A discrete event simulation, an optimization model, and a forecasting method were integrated into the system's architecture to analyze performance when facing different scenarios with several changes in the system's characteristics. We conducted a survey on energy trading, considering that renewable energy generators were incorporated into the algorithm to determine the interaction between data centers and smart grids. The experimental results demonstrate that the proposed system has great potential in improving energy efficiency under different operating conditions in the data centers.

Keywords: Energy efficiency, resource management, data center, smart grid, renewable energy

1.0 INTRODUCTION

Cloud computing environments are based on the performance of several large-scale data centers that simultaneously provide different kinds of services to millions of customers. It has become a necessity in conducting our daily activities, for instance, services, like commerce, education, business, social networking, communication services, and others are related to data centers. Due to the use of Internet-based services, the information technology (IT) infrastructure is a demanding topic. As a result, managing data centers has become an important issue involving several areas, such as energy consumption and production, IT performance, and cooling [1].

In recent years, the number of data centers has grown considerably; likewise, data capacity has been modified to constantly handle new requirements, for example, Google quadrupled the number of servers in the 2000-2010 period to support the operation of Google searches, YouTube viewing, Gmail messaging [2]. Nonetheless, this trend negatively affects the environment; it has generated a high carbon footprint derived from CO2 emissions produced by data centers. These environmental concerns have triggered green initiatives worldwide, one of the sustainable solutions that apply to this case is the integration of renewable energy to reduce energy dependency on the grid and emissions. Currently, wind and solar power are the sources to energize small and medium data centers

[3], even big companies are pointing in this direction; for instance, Apple and McGraw-Hill have implemented solar arrays for some of their data centers recently built. On the other hand, Google is taking advantage of the geographic location of its data center in Hamina-Finland, which uses a cooling system based on seawater supplied from the bay.

Similarly, high energy consumption by data centers concerns IT enterprises and governments. It is estimated that data centers consumed approximately 1.5% of all electricity worldwide in 2011 [4] because of continuous operation throughout the year. In fact, a cooling system can consume up to 50% of total energy [5]. It is also important to consider that most data centers have been designed based on the worst-case scenario and so one consequence is the oversize of the components that make them inefficient systems with high operating costs. In the past decade, power consumption of data centers worldwide rose from 70 billion to 330 billion kWh and this value is projected to increase to more than triple by 2020 [6], which represents a significant economic investment for different companies in this sector and demonstrates the importance of efficiency in the operation of data centers.

There are two methods to reducing energy consumption in data centers. The first method is to relocate the data centers to areas with cold climate, so thousands of servers can be cooled by using a smaller amount of cooling equipment. For example, Facebook recently built a data center at the edge of the Arctic Circle in Sweden, which uses outside air for cooling instead of air conditioning [2]. The second method consists in developing algorithms that operates within the data centers to allow servers to improve their performance by reducing power consumption [7].

Modern data centers must be eco-friendly and manage energy consumption from technical and economic perspectives. As a new trend has been reported the use of green data centers which work on a smart grid environment [1]. A smart grid uses an electrical grid including energy-efficient resources, and renewable energy resources (mainly wind power and solar power). Renewable energy generation is highly variable and needs sophisticated control systems to achieve reliability and efficiency. On the other hand, the workloads of large-scale data centers are also variable, we believe that coordinated resource management and energy management approach could help data centers to use renewable energy more effectively.

This paper presents an integrated set of discrete-event simulation, nonlinear optimization, and forecasting used to analyze several aspects involving data center integration with smart grids environments. For the resource allocation algorithm, the principle of operation is similar to one reported in [8, 9]. The aspect of power consumption caused by IT and cooling infrastructure was evaluated due to the great importance of data center management, as discussed in [10, 11]; therefore, two scenarios were created in two different geographic locations with different atmospheric conditions, to determine how to affect power consumption in data centers. Integration of data centers and smart grids was also considered in the algorithm, incorporating the use of renewable energy generators [12, 13], by focusing on energy trading with the grid. The results showed a trade-off between service level agreement (SLA) fulfillment, turning on servers, and selling generated energy. This trade-off depends on parameters, like energy selling prices, generator capacity, and power consumption of the cooling infrastructure.

There is a great potential in considering integrated management of IT systems, cooling, and energy in data centers. Creating this integrated solution is the aim of this paper. Particularly, the solution focuses on optimizing the workload management by integrating the supply of renewable energy, dynamic prices of renewable and non-renewable energy, and the supply of cooling to improve the power efficiency of the data center. An important aspect of our work is the change of the demand and the allocation of IT resources within the data center, considering the generation and use of renewable and nonrenewable energy according to its availability and cooling efficiency. A contribution considered in this work is the addition of a detailed cost model to the optimization problem contemplating the use of nonrenewable energy. To validate our model, several experiments were conducted to highlight the practicality of the approach used.

2.0 METHODOLOGY

The proposed model was developed on an application-based architecture with multiple tiers, where three types of servers were allocated in clusters, to respond to specific requests from different kinds of clients [14]. Hence, the data center model used was structured in three clusters, where servers performed various tasks (the web, database, and application services), as shown in Figure 1.

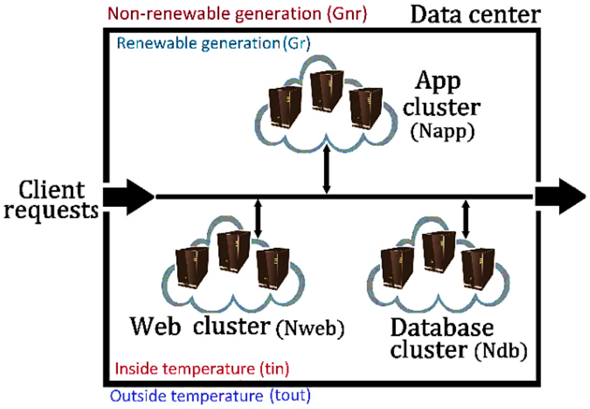


Figure 1 Data center model based on an application multi-tier architecture including web, database and application services

The simulation model included power generation for the data center and a cooling method. The cooling method considered in this work, referred to as outside air temperature (OAT) cooling [15], uses the difference between the inside temperature t_{in} of the data center and the outside temperature t_{out} of the place where the data center is located. In OAT cooling, outside cold air is directed into the data center using air-side economizer to cool down servers [16], so the data center uses OAT cooling when $t_{out} \leq t_{in}$. According to this method, when the outside air temperature decreases, high energy efficiency of the cooling system is obtained, resulting in lower energy consumption related to cooling. Other cooling methods exist that contemplate distinct variations between the temperature inside and outside the data center [17]. Regarding power generation, the model included renewable and non-renewable generation, in order to decide on the best selling price for both. The model was implemented in a discrete-event simulation, which was used to evaluate the results obtained from the optimization algorithm and to achieve more realistic results than the ones given by the algorithm.

The model depicted in Figure 1 was analyzed as a queuing network with a total number of Ns servers, where a dispatcher was located at the entrance of the data center, assigning each request to its corresponding cluster. As soon as the request leaves the cluster, it could go to another cluster, or it could leave the data center. These decisions depend on routing probabilities among clusters. Figure 2 shows the architecture of the simulation model.

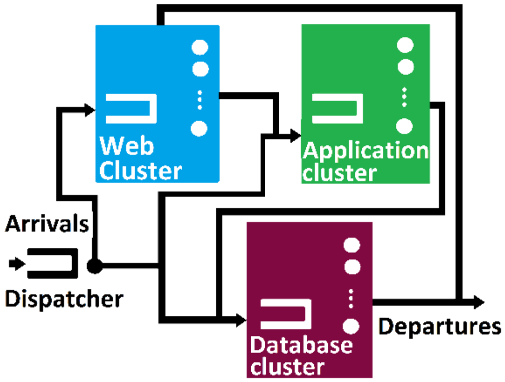


Figure 2 Data center model used in the simulation

Dispatcher performance (assuming inter-arrival and service times exponentially distributed) was considered as a $M/M/1$ queue and each cluster was designated as a $M/M/k$ queue with $k = Nweb, Napp$, or Ndb, where $Nweb$ is the number of servers in the web cluster, $Napp$ is the number of servers in the application cluster, and Ndb is the number of servers in the database cluster. The simulation was conducted according to the model explained above with support from the stochastic simulation in Java (SSJ) library [18].

To achieve an actual approximation of the dynamics of a data center, a real trace was analyzed to obtain arrival and service rates. Full information on a particular day, containing approximately 15.000 logs during the day, was taken and separated at 2-hour intervals [19]. Table 1 presents these variations of arrival and service rates during the day considered.

Table 1 Arrival and service request rate during a particular day in the studied trace

Hour (Interval)	Arrival rate (requests/min)	Service rate (requests/min)
0-2	8.40	0.153
2-4	10.90	0.151
4-6	9.31	0.190
6-8	8.87	0.210
8-10	10.68	0.228
10-12	8.25	0.196
12-14	8.89	0.200
14-16	8.98	0.180
16-18	9.03	0.183
18-20	9.97	0.209
20-22	8.62	0.192
22-24	9.68	0.200
Minimum	8.25	0.151
Maximum	10.90	0.228

An optimization model that integrates data center performance and some features involving a smart grid environment was developed. This model can be executed dynamically, meaning that in each interval of time, the results are updated according to the variations of the problem inputs. Because of this, time was divided into T intervals and the optimization problem was defined for each interval t of time $(1 \leq t \leq T)$. Hence, three aspects were considered in the optimization model: First, it was the SLA fulfillment between the data center and several types of clients with day. Second, the power consumption in the data center was considered, which was separated between server power consumption and cooling power consumption. The last consideration was the different types of power generation in the data center and the transactions that can be done with this energy.

The whole system consisted of three main modules. The forecaster was used to predict the behavior of several parameters involved in the optimization model. Then, the optimization algorithm was executed to obtain the optimal resource allocation of the data center for the new interval. Additionally, the optimization algorithm achieves the best energy management configuration for the new interval. The simulation was used to evaluate the performance of the data center specifically in terms of average response time per type of client, using the optimal server allocation. This process is depicted in Figure 3.

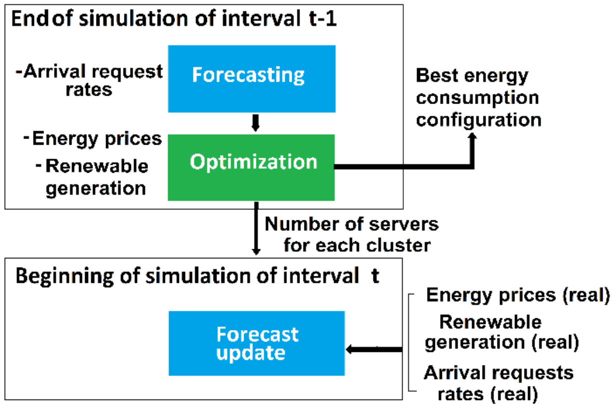

End of simulation of interval t-1
-Arrival request rates
Forecasting
-Energy prices
-Renewable generation
Optimization
Best energy consumption configuration
Number of servers for each cluster
Beginning of simulation of interval t
Forecast update
Energy prices (real)
Renewable generation (real)
Arrival requests rates (real)

Figure 3 Module integration when a time interval ends, and the other begins

2.1 SLAs Fulfillment

The performance metric evaluated in this model was the average response time in the data center for each type of client. Assuming Nc types of clients that entered a service level agreement ($1 \leq i \leq Nc$), each was assigned a cost function that depends on the average response time. Additionally, the average response time depends on the number of servers allocated to each cluster. The expression to be used to calculate the average response time for multi-server queues is not closed, which adds to a high level of complexity to find solutions to the objective function of the optimization problem. Hence, the following simplification was performed in the optimization model: each cluster was designated as a group of independent servers ($M/M/1$ queues) with the same arrival rate, which is the effective arrival rate to the cluster divided by the number of servers assigned to the cluster. According to this, the average number of requests in each cluster (for the type of client i) will be the sum of the mean number of requests in each server (in each $M/M/1$ queue) as shown in Equation (1).

$$L_{k,i} = N_k \frac{\rho_k}{1-\rho_k}, \quad where\ \rho_k = \frac{\lambda_{k,i}/N_k}{\mu}, \quad k\ \in \{web, app, db\} \qquad (1)$$

The average number of requests in the dispatcher ($L_{disp,i}$) was calculated in the same way as for one server of one cluster (because the dispatcher is an $M/M/1$ queue) as:

$$L_{disp,i} = \frac{\rho_{disp}}{1-\rho_{disp}}, \quad \rho_{disp} = \frac{\lambda_i}{\mu_{disp}} \qquad (2)$$

In the Equations (1), (2), μ is the server service rate for a request (assuming $\mu_{disp} \gg \mu$) and each arrival rate is the effective arrival rate to each station, calculated assuming that the system itself is a Jackson network. A Jackson network is a set of queues (now named stations) with different routing probabilities, given by a route matrix P (each type of client has a different matrix) with the form,

$$P = \begin{bmatrix} 0 & P_{disp,web} & P_{disp,app} & P_{disp,db} \\ 0 & 0 & P_{web,app} & P_{web,db} \\ 0 & P_{app,web} & 0 & P_{app,db} \\ 0 & P_{db,web} & P_{db,app} & 0 \end{bmatrix}$$

This matrix was used to obtain the effective arrival rates to each station. By using these arrival rates, it was possible to collect performance metrics for each station and, therefore, obtaining performance metrics for the whole network. The vector $\hat{\lambda}$ containing the effective arrival rates to each station was computed, as shown in Equation (3).

$$\hat{\lambda} = \lambda(I-P)^{-1} \qquad (3)$$

I corresponds to the identity matrix. The vector λ contains the external arrival rates to each station. For this model, this vector has zero values except in its first position, which is the arrival requests rate to the data center. By using the effective arrival rates to each station, it is possible to calculate the average number of requests for each station, while taking into account the specific expressions for each queuing system. The average number of request in the system L_i was computed by adding each value of the mean number of requests in each station, as shown in Equation (4).

$$L_i = L_{disp,i} + L_{web,i} + L_{app,i} + L_{db,i} \qquad (4)$$

Finally, the average response time in the system ($T_{av,i}$) was calculated by using Little's law, which divides the average number of requests in the system by the arrival requests rate to the network, as shown in Equation (5).

$$T_{av,i} = \frac{L_i}{\lambda_i} \qquad (5)$$

There are Nc different average response times, all depending on three decision variables: the number of servers in the web cluster ($Nweb$), the number of servers in the application cluster ($Napp$) and the number of servers in the database cluster (Ndb). Similarly, each average response time has a cost function associated. For a type of client i, the cost function was performed as a linear function with a positive slope as average response time increases [8].

If the data center performance can guarantee that the average response time for a type of client i in the interval t is lower than the critical time T_c, there will be no penalty incurred and the data center will receive a benefit b_i, as shown in Equation (6).

However, if the average response time exceeds the critical time, the data center must pay a penalty that increases proportionally with the penalty slope m_i.

$$u_{i,t} = \begin{cases} -b_i & if\ T_{av,i} < T_c \\ m_i(T_{av,i} - T_c) & if\ T_{av,i} \geq T_c \end{cases} . \tag{6}$$

2.2 Power Consumption

As stated above, the power consumption in the data center mainly depends on how many servers are working in each interval of time and how much energy it takes to cool down these devices. Assuming that each server consumes E_{work} amount of energy in one-time interval, the server's power consumption is defined as:

$$E_{c,t} = (N_{web,t} + N_{app,t} + N_{db,t})E_{work} \tag{7}$$

Where $N_{web,t}, N_{app,t}, N_{db,t}$ corresponds to web, application and database servers working in each interval of time.

The power consumption of the cooling devices is directly related to the power consumption of IT devices. However, this relationship can change depending on the type of cooling process used in the facility. An outside air cooling method was implemented in the data center [11], with power consumption represented, as shown in Equation (8).

$$E_{cool,t} = kE_{c,t}^3 \tag{8}$$

Parameter k is proportional to the difference between the outside temperature and the inside temperature in the facility,

$$k = \alpha\left(\frac{1}{t_{in} - t_{out}}\right) \tag{9}$$

The inside temperature was established at 35°C, whereas the outside temperature depends on the geographic location of the data center. The parameter α when the temperature is 35°C is approximately 0.03 given by [11]. Another important issue was to capture the dynamics between turning servers on and off for every consecutive interval. The process of turning on a server refers to the change from an idle state to an active state, the inverse process is known as turning off the server. To accomplish this goal, new decision variables were defined: the number of servers turned on N_{on} during the interval t and the number of servers turned off N_{off} related to the number of servers working N_{work} on interval t by Equation (10).

$$N_{work,t} = N_{work,t-1} + N_{on,t} - N_{off,t} \tag{10}$$

During each time interval, some servers are turned on and off according to the tradeoff proposed in the optimization problem and complying with the restrictions imposed. The on process makes the server have an extra consumption of energy in the interval t. Additional energy consumption occurs at the start of each interval; however, when using a discrete system, an increase should be considered during the whole interval. Assuming that a server consumes E_{on} amount of energy when turned on, the power consumption caused by N_{on} servers turned on was calculated based on Equation (11).

$$D_{on,t} = N_{on,t}E_{on,t} \tag{11}$$

Finally, the total power consumption of the system during a time interval was defined as the sum of the three quantities considered, as shown in Equation (12).

$$E_{total,t} = E_{c,t} + E_{cool,t} + D_{on,t} \tag{12}$$

2.3 Energy Generation

This model involved two different types of power generation: renewable and non-renewable. This separation was performed because of different processes and costs associated with each type of power generation, which allows determining the best selling price of power generated by both methods. Because the data center uses renewable and nonrenewable energy for its operation, during an interval t, the use of one or both types of energy can be lower than the installed generation capacity. Variables required to define the optimization problem are listed ahead (at interval time t):

$G_{r,t}$: Renewable energy generated.
$P_{r,t}$: Selling price of renewable energy
$P_{nr,t}$: Selling price of non-renewable energy
$C_{grid,t}$: Grid energy cost.
$C_{nr,t}$: Non-renewable generation capacity.

The optimization program included the decision variables listed as follows:

$G_{rs,t}$: Renewable energy sold.
$G_{nrs,t}$: Non-renewable energy sold.
$G_{nr,t}$: Non-renewable energy not sold.
$x_{grid,t}$: Energy taken from the grid.

The following optimization problem was formulated to guide the proposed non-linear program to manage the resource allocation of an energy-efficient data center in a smart grid environment. We focus on minimizing power consumption to reduce the environmental impact while satisfying the clients' SLAs.

$$\min \sum_{i=1}^{Nc} u_{i,t}\left(T_{av,i}\right) + C_{grid,t}\, x_{grid,t} - G_{rs,t}P_{r,t} - G_{nrs,t}P_{nr,t}$$

$$\begin{aligned}
&s.t\ N_{work,t} \leq N_s\\
&N_{work,t} = N_{work,t-1} + N_{on,t} - N_{off,t}\\
&N_{work,t} = N_{web,t} + N_{app,t} - N_{db,t}\\
&G_{nr,t} + \left(G_{r,t} - G_{rs,t}\right) + x_{grid,t} = E_{total,t}\\
&G_{rs,t} \leq G_{r,t}\\
&x_{grid,t},\ G_{rs,t}, G_{nrs,t}, G_{nr,t}, N_{off,t} \geq 0\\
&N_{on,t}, N_{web,t}, N_{app,t}, N_{db,t} \geq 0\\
&N_{off,t}, N_{on,t}, N_{web,t}, N_{app,t}, N_{db,t} \in\ N_s\\
&G_{nrs,t} + G_{nr,t} \leq C_{nr,t}
\end{aligned}$$

This optimization problem is nonlinear because of the expression for the average response time and the expression of the cooling method consumption. The objective function focused on the costs incurred in the data center, both for performance and power consumption. Also, in case the energy generated could be sold the possible benefits were included.

2.4 Forecasting of Time Series

Several parameters needed to be forecasted upon solving the optimization problem, given their actual values are only known during the corresponding time interval, but the planning decisions must be addressed in advance. These included the energy grid price, the renewable energy generated, and the generated energy sale price. According to this, it was necessary to use a method for these time series to predict their behavior for the next interval of time. The exponential smoothing method was implemented, which is a widely used forecasting technique based on assigning weights to different observations of the time series Y_t, where the newer observations get more weight than the older ones [20]. When the time series presents a seasonality behavior (with a seasonality period of duration s), the method learns about the prediction by using this feature of the series. To accomplish this forecasting process with seasonality behavior, the method uses two parameters: L_t and S_t, which were computed by using Equation (13), (14), as stated in [21] where $0 \leq \alpha, \gamma \leq 1$,

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)L_{t-1} \qquad (13)$$

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s} \qquad (14)$$

The prediction F_{t+1} for the next time interval is expressed regarding these two parameters as,

$$F_{t+1} = L_t * S_{t-s+1} \qquad (15)$$

It is important to mention that, depending on the values of α and γ, it is possible to obtain a better prediction of the time series. These values should be chosen to minimize the mean squared error [22] to achieve more realistic results when the optimization is executed. Finally, a good way to initialize the first s values of S and the value L_s is shown in Equation (16), (17).

$$L_s = \frac{1}{s}(Y_1 + Y_2 + \cdots + Y_s) \qquad (16)$$

$$S_k = \frac{Y_k}{L_s}, \quad k = 1, 2, \ldots, s \qquad (17)$$

The first s values of the time series Y_t need to be known ahead of time to initialize the forecasting parameters.

3.0 RESULTS AND DISCUSSION

To evaluate the performance of the proposed system, a base scenario was created with four types of clients and four routing matrices, which were chosen randomly:

$$P_1 = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0.1 & 0 & 0.1 \\ 0 & 0 & 0.5 & 0 \end{bmatrix} \quad P_2 = \begin{bmatrix} 0 & 0.6 & 0.4 & 0 \\ 0 & 0 & 0 & 0.9 \\ 0 & 0.1 & 0 & 0.1 \\ 0 & 0 & 0.5 & 0 \end{bmatrix}$$

$$P_3 = \begin{bmatrix} 0 & 0.3 & 0.3 & 0.4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \end{bmatrix} \quad P_4 = \begin{bmatrix} 0 & 0.2 & 0.6 & 0.2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Four clients were chosen considering that the number is enough to note the behavior of different effective arrival rates within the system. However, the proposed system has the capacity of being scalable and it may work with a higher number of clients. There is no limit to the number of clients that the system can hold, nonetheless, the models of queues might get saturated causing inaccuracies on the results of the optimization model. An appropriate IT equipment would allow working with a higher number of clients.

The cost function parameters for each type of client are shown in Table 2. These parameters were selected to assign the same weight to SLA fulfillment and power consumption costs. The currency used in costs was the Colombian peso (COP), to be consistent with the geographical location of the scenarios. Also, the base scenario included the following features:

- Dispatcher service rate: 6000 requests/min.
- Server service rate: 0.228 requests/min.
- Total number of servers: 500.

Table 2 Cost function parameters for each type of client

Type of Client	b (COP)	m(COP/s)	$T_c(s)$
1	2000	2000	520
2	3000	1000	530
3	1000	5000	228
4	5000	1000	200

Both service and arrival rates (for each type of client) were taken according to the results obtained for the Google trace analyzed [19]. Regarding energy generation, the following parameters were considered in the base scenario:

- Non-renewable generator: 1000 kW diesel generator.
- Renewable generation: 1000 kW solar array and 850 kW wind generator. We assumed the generator is located in the city of Barranquilla-Colombia, specifically in Las Flores. This assumption was made to provide accurate data on wind speed because the power generated by a wind generator is directly proportional to wind speed.
- Purchase price of electricity (market price): taken from April 4 to April 7 (2017) with hourly resolution [23].
- Renewable energy selling price: 12.5% less than the purchase price of energy.
- Non-renewable energy sale price: 6.25% less than the purchase price of energy.

The reduction percentages of the selling price of generated energy were assigned based on the information by [24]. However, given that operation and maintenance costs of renewable energy are higher, the reduction of its selling price is bigger.

The parameters related to power consumption were as follows:

- Energy consumption of a server when it is switched on: 0.45 kWh. Energy consumption of a server working: 0.4 kWh [25]. The difference in energy consumption between a server that has just been turned on and one that was already operating is of 0.05 kWh.
- Difference between external and internal temperature: 7°C (assuming an average external temperature of 28°C); this value is necessary to obtain the energy consumption of the cooling infrastructure.

The system was tested by simulating a complete day divided into 1-hour intervals. In every interval, the forecaster was used to predict energy prices, renewable generation, and client arrival rates; thereafter, these predictions were used to execute the optimization algorithm (at every interval) that determines the number of servers to assign to each cluster and the values of the other decision variables.

Figure 4 shows the cost of the objective function during the day. There is a difference between the results of the optimization algorithm and the simulation, which was caused by two aspects. First, the simplification of the calculation of the average response time is always an upper bound on the actual average response time given by an $M/M/k$ queue. This statement is true because the main assumption for the simplified model is that servers belonging to a cluster divide their workload equally and work independently. This is not the case of a $M/M/k$ queue where the system dynamically distributes the workload among servers, making them better exploit resources. The second aspect is related to the forecasting method because the forecaster cannot predict with 100% accuracy; thus, there will be minor errors in the optimization algorithm compared to the real data of the time series used to obtain the costs of the objective function in the simulation. The mean absolute error was obtained for the forecasting method, and it always was lower than 13.5%.

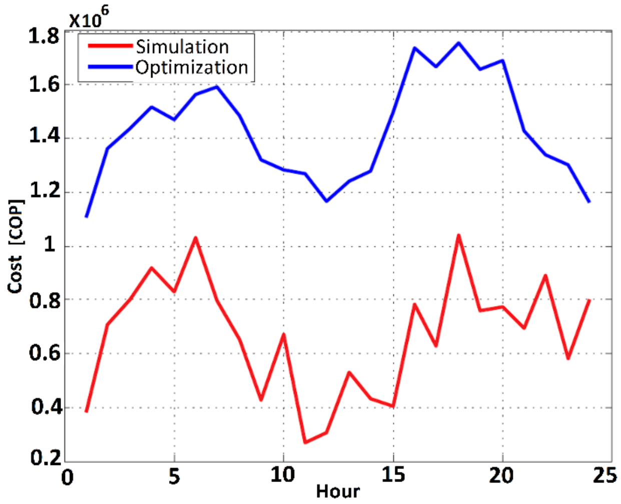


Figure 4 Objective function costs for the optimization algorithm and simulation during the day of simulation

The number of servers allocated to each cluster during the day is depicted in Figure 5. It is worth mentioning that the optimization algorithm does not turn on all servers at any hour of the day, as shown in Figure 5, because it generates a higher costs due to power consumption, even though the data center has to assume an SLA violation penalty. Likewise, by having a number of servers turned off, the data center obtains a surplus of generated energy, which can be sold. The optimization algorithm chooses to sell non-renewable energy because it is more profitable (its selling price is greater than the renewable energy sale price).

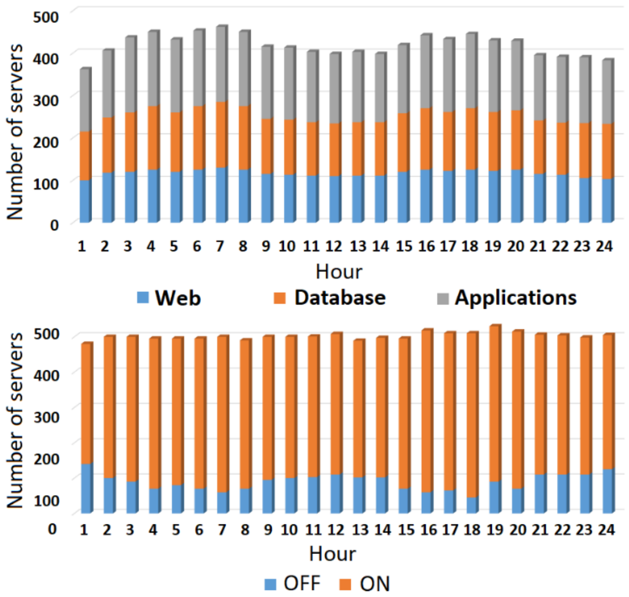


Figure 5 (a) Number of servers allocated to each cluster, (b) Number of servers turned on and turned off

Figures 6, 7, and 8 show the differences between the values predicted and the actual observations for all the forecasted time series. The values of α and γ were chosen to be the same for all series (0.5) [26]. In effect, the performance of the forecasting method is adequate to obtain well-predicted data. However, this performance could be damaged if the time series presents abrupt changes during the day. To improve the performance of the forecasting method, it is possible to assign the values of α and β that minimize the mean squared error in each of the time series used.

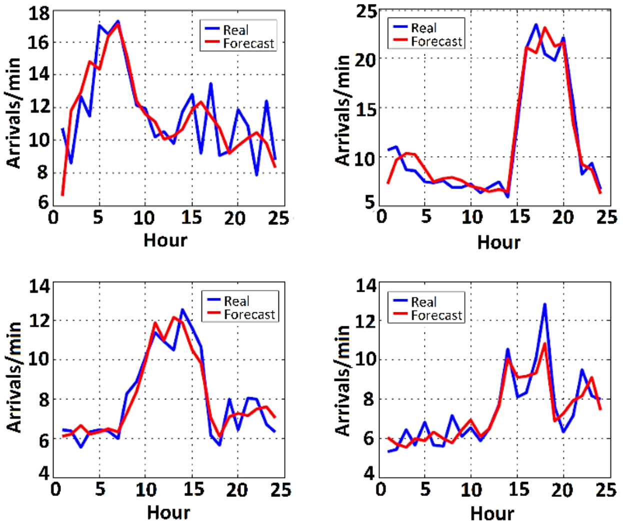


Figure 6 Comparison between the real time series of the arrival requests rate for each type of client and the results of its prediction

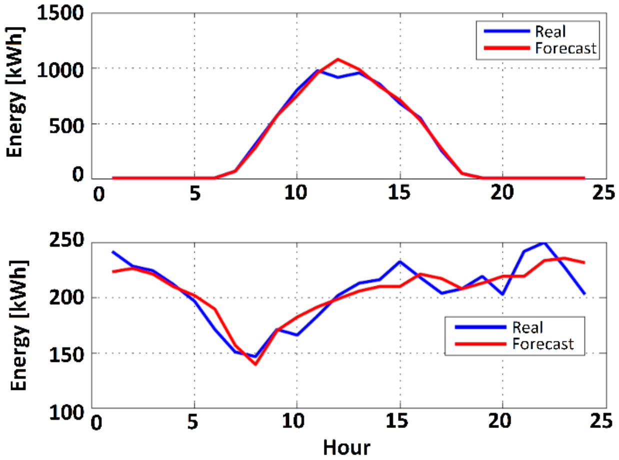


Figure 7 (a) Comparison between the real-time series and its prediction for the renewable energy generated using solar and (b) wind generators

Several experiments were conducted to evaluate data center performance when encountering changes in various parameters.

Experiment 1: This experiment changed the geographic location of the data center to Bogotá-Colombia because the outside temperature is lower in Bogotá than in Barranquilla, the OAT cooling method would be more efficient regarding power consumption. If the power consumed by the cooling infrastructure is reduced, it could be possible to turn on more servers because, as seen in the base scenario, surplus energy was generated. If this surplus is used to provide energy to new servers, there will be no grid power cost.

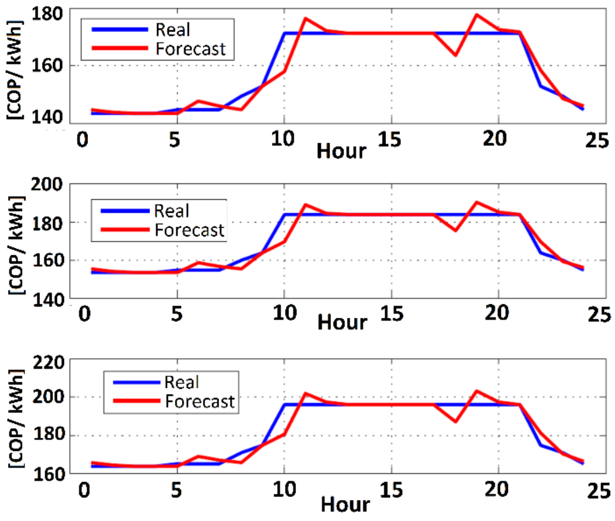


Figure 8 (a) Comparison between the real-time series and its prediction for the renewable energy selling price, (b) the non-renewable energy selling price, and (c) the grid energy cost

In fact, Figure 9 shows the new resource allocation for this scenario, indicating that most of the time all the servers are on. Also, if the geographic location is changed, there is a change in renewable energy generated. In this experiment, wind and solar capacity were reduced to consider the geographic change. Figure 10 shows the total reduction of power taken from the grid compared to the base scenario. The peak energy values for both scenarios have differences of almost 50%; this suggests that the geographic location change saves nearly half the energy taken from the grid when the data center is using an OAT cooling method.

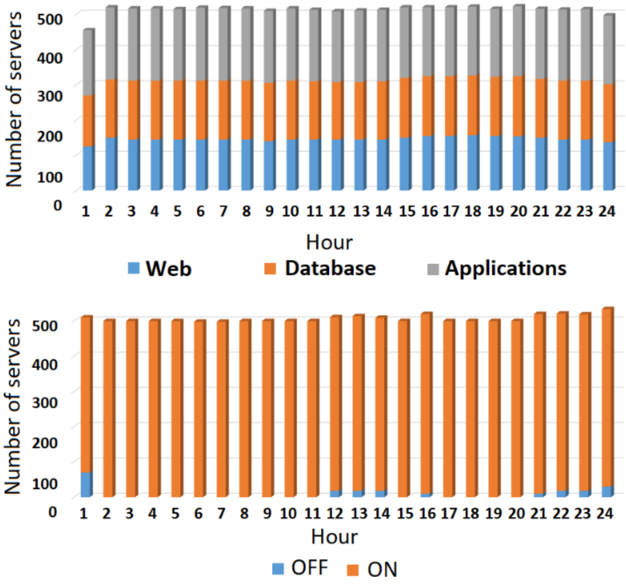


Figure 9 (a) Number of servers allocated in each cluster, (b) Number of servers turned on and turned off, when the data center location is changed to Bogotá

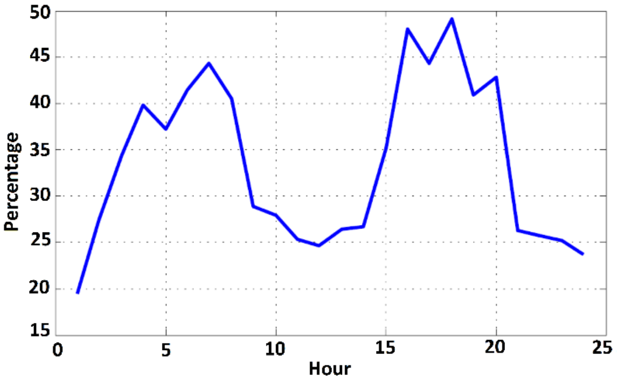


Figure 10 Reduction of the energy taken from the grid when the data center location changes

Experiment 2: We reduced by half the capacity of all renewable and non-renewable generators to determine if the optimization algorithm would change its resource allocation policy. In effect, the number of servers turned on was less than the results from Figure 9, while quite similar to those shown in Figure 5. The number of servers turned on depends on the installed capacity of generators.

Experiment 3: A sensitivity analysis on the energy selling price of renewable and non-renewable energy was conducted. We did 16 runs of the scenario with different values of this price based on a reduction percentage of the energy grid cost. Table 3 shows the results of this sensitivity analysis, where each cell is the average benefit on the day when the renewable energy selling price, P_r, and the non-renewable energy sale price, P_{nr}, are reduced by those percentages.

Table 3 Average day benefit when selling generated energy. The energy selling price is chosen by reducing the power grid costs by the percentages above

		P_{nr}	P_{nr}	P_{nr}	P_{nr}
		25%	**12.5%**	**4%**	**1%**
P_r	25%	\$8330.23	\$9750.2	\$12090.3	\$13341.01
P_r	12.5%	\$9431.8	\$9750.2	\$12090.3	\$13341.01
P_r	4%	\$11014.4	\$11014.4	\$12090.3	\$13341.01
P_r	1%	\$13216.8	\$13216.8	\$13478.5	\$14583.9

As a result, the lower the gap between the selling price and the energy grid cost, the better the benefit because it becomes more attractive to sell energy; this is why the best energy sale price for both types of generation is 1% above the energy cost. With these results, the data center could aim to reduce energy production costs to obtain better profits when energy is sold.

Experiment 4: Another aspect evaluated was the effect of abrupt changes in the forecasted data. Figure 11 shows this effect on the arrival requests rate for client 2. Indeed, the forecast fails to accurately predict some points of the day and this makes the optimization algorithm provide a solution that is not necessarily the best. In fact, the average cost of the objective function during the day is 20.83% greater in this case than with the new geographic location, this increased cost is caused because the forecaster predicts lower arrival rates for client 2, causing the optimization model to assign a lower number of servers and increasing penalties in SLAs.

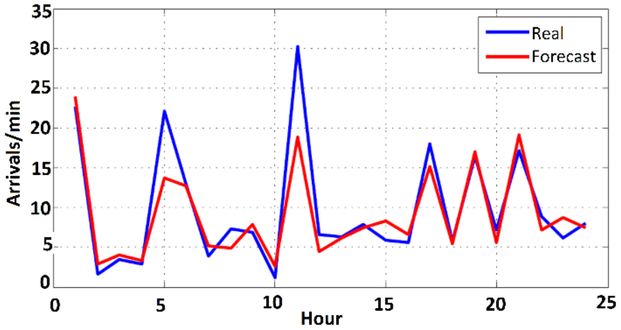


Figure 11 Comparison between the real-time series and its prediction for the arrival request rate of client 2 when there are abrupt variations in the data

Experiment 5: The average solving time of the optimization algorithm was estimated as a function of the data center size. By increasing the total number of servers in the data center, we also increased the SLA penalties and the input demand of each type of client to keep the same order of magnitude. As shown in Figure 12, the average solving time increases with the number of servers. However, this increase only gets a big difference when the data center has more than 104 servers. If we assume that each interval the algorithm requires this average time to solve itself, the simulation will take approximately T times this average time to complete. Thus, for a data center of 105 servers, results will be given in approximately 1.5 min (without taking into account simulation and forecasting). The total execution time of the system (including forecasting and simulation) with our configuration of clients and servers is approximately 12 minutes, which is faster than the one reported by [8], that exceeds 30 minutes.

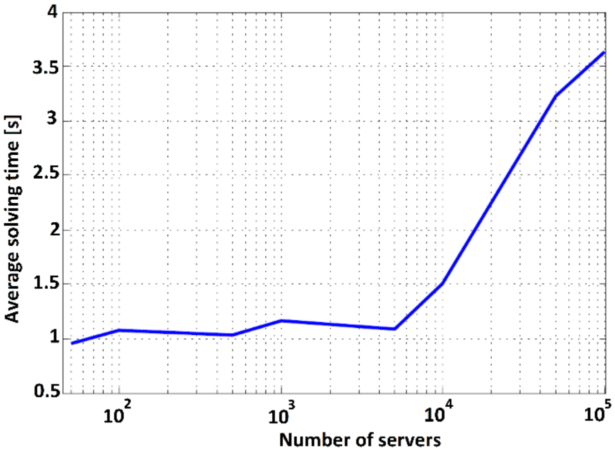


Figure 12 Average solving time of the optimization algorithm in function of the data center size

4.0 CONCLUSION

This work presents an integrated workload management system for data centers constrained by service level agreements, power consumption, and energy generation. Additionally, other factors were considered in the integration between data centers and smart grids. Such as electricity price, cooling, and the availability of renewable energy. The proposed system simulates the performance of the data center using a forecasting method, and an optimization model that involves all the variables mentioned before. To ensure an actual approximation of the dynamics of a data center, a real trace was analyzed to obtain arrival and service times.

A base scenario was developed and used to study the effects of different operating conditions on the system's performance; the behavior of energy trading with the grid was also determined from these experiments. An OAT cooling method was implemented in the facility; thus, allowing to determine that the geographic location is critical in the total power consumption, particularly in cold places where it generates the possibility of increasing considerably the number of servers turned on. When the data center location was changed to Bogotá, the total energy consumption from the grid where almost 50% (peak value) less than the base scenario (with the location in Barranquilla). Because of this reduction, the number of servers turned on increased to almost the total number of servers in the place. Although the data center could turn on more servers (between 10-15 hours) since there was a surplus of generated energy, the optimization algorithm decided to sell this energy using the best-selling price, which was for non-renewable energy. Also, a trade-off was noticed between power consumption and service level agreements, even though the data center had the possibility of keeping all the servers on, it kept some of them off.

A sensitivity analysis was applied to the selling prices of energy generated; the results showed that the highest average benefit during a day was achieved when there was a small difference between the selling price and the energy grid cost. This would lead the data center to implement policies to reduce operating costs of generation to obtain higher utility.

The results obtained from the optimization model and the simulation presented differences due mainly to the simplifications made to the average response time calculation in the optimization algorithm; thus, the simulation provided results close to reality regarding costs for the objective function. The forecasting method presented an odd behavior when the input time series had abrupt changes because it calculated a lower value than the actual data, which causes increased SLA penalties.

A limitation of this work is the calculation of the average response time in the optimization model because it is not done independently of the probability density of the parameters. This could be achieved by generating a function based on a multivariable regression because the average response time depends on the number of servers in each cluster and the arrival rates of each type of client. In terms of the simulation model, it could include other aspects of the data center dynamics such as server virtualization, and taking advantage of its discrete-event condition, it could consider power grid failures and the response of the data center generators.

Finally, the results of the study demonstrate that the proposed system has a great potential for improving energy efficiency under different operating conditions in the data centers. Future work complementing the study presented in this paper consists of contemplating different cooling schemes, and on analyzing the impact of workload management in the size of renewable and IT infrastructure.

Acknowledgement

This work is the results of a research project funded by Universidad ECCI through the internal call N° 4 of 2017-2018.

References

[1] X. Wang, Z. Du, Y. Chen, M. Yang. 2015. A Green-aware Virtual Machine Migration Strategy for Sustainable Datacenter Powered by Renewable Energy. *Simulation Modelling Practice and Theory*. 58(Part 1): 3-14. Doi:10.1016/j.simpat.2015.01.005.

[2] M. Pedram. 2012. Energy-efficient Datacenters. *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems.* 31(10): 1465-1484. Doi:10.1109/TCAD.2012.2212898.

[3] Í. Goiri, M. E. Haque, K. Le, R. Beauchea, T. D. Nguyen, J. Guitart, J. Torres, R. Bianchini. 2015. Matching Renewable Energy Supply and Demand in Green Datacenters. *Ad Hoc Networks.* 25: 520-534. Doi:http://dx.doi.org/10.1016/j.adhoc.2014.11.012.

[4] A. Wierman, Z. Liu, I. Liu, H. Mohsenian-Rad. 2014. Opportunities and Challenges for Data Center Demand Response. *Green Computing Conference (IGCC), 2014 International.* 1-10. Doi:10.1109/IGCC.2014.7039172.

[5] E. Oró, V. Depoorter, A. Garcia, J. Salom. 2015. Energy Efficiency and Renewable Energy Integration in Datacentres. Strategies and Modelling Review. *Renewable and Sustainable Energy Reviews.* 42: 429-445. Doi: 10.1016/j.rser.2014.10.035.

[6] T. Mastelic, I. Brandic. 2015. Recent Trends in Energy-efficient Cloud Computing. *IEEE Cloud Computing.* 2(1): 40-7. Doi:10.1109/MCC.2015.15.

[7] D. Sitaram, H. L. Phalachandra, G. S, S. H. V, S. TP. 2015. Energy Efficient Data Center Management Under Availability Constraints. *Systems Conference (SysCon), 9th Annual IEEE International, 2015.* 377-38. Doi:10.1109/SYSCON.2015.7116780.

[8] D. Ardagna, M. Trubian, L. Zhang. 2017. SLA based Resource Allocation Policies in Autonomic Environments, *Journal of Parallel and Distributed Computing*. 67(3): 259-270. Doi:http://dx.doi.org/10.1016/j.jpdc.2006.10.006.

[9] Q. Zhang, M. F. Zhani, S. Zhang, Q. Zhu, R. Boutaba, J. L. Hellerstein. 2012. Dynamic Energy-aware Capacity Provisioning for Cloud Computing Environments. *Proceedings of the 9th International Conference on Autonomic Computing, ICAC '12.* 145-154. Doi:10.1145/2371536.2371562.

[10] Q. Tang, S. K. S. Gupta, G. Varsamopoulos. 2008. Energy-efficient Thermal-aware Task Scheduling for Homogeneous High Performance Computing Data Centers: A Cyber-physical Approach. *IEEE Transactions on Parallel and Distributed Systems.* 19(11): 1458-1472. Doi:10.1109/TPDS.2008.111.

[11] R. Das, J. O. Kephart, J. Lenchner, H. Hamann. 2010. Utility-function-driven Energy-efficient Cooling in Data Centers. *Proceedings of the 7th International Conference on Autonomic Computing, ICAC '10, ACM, New York, NY, USA*. Doi:10.1145/1809049.1809058.

[12] D. Gmach, J. Rolia, C. Bash, Y. Chen, T. Christian, A. Shah, R. Sharma, Z. Wang. 2010. Capacity Planning and Power Management to Exploit Sustainable Energy. *2010 International Conference on Network and Service Management*. 96-103. Doi:10.1109/CNSM.2010.5691329.

[13] A. H. Mohsenian-Rad, A. Leon-Garcia. 2010. Coordination of Cloud Computing and Smart Power Grids. Smart Grid Communications (SmartGridComm). *2010 First IEEE International Conference on*. 368-372. Doi:10.1109/SMARTGRID.2010.5622069.

[14] M. Arregoces, M. Portolani. 2003. *Data Center Fundamentals.* Cisco Press.

[15] H. Zhang, S. Shao, H. Xu, H. Zou, C. Tian. 2014. Free Cooling of Data Centers: A Review. *Renewable and Sustainable Energy Reviews.* 35(0): 171-182. Doi:10.1016/j.rser.2014.04.017.

[16] M. Islam, S. Ren, N. Pissinou, A. Mahmud, A. Vasilakos. 2015. Distributed Temperature-aware Resource Management in Virtualized Data Center. *Sustainable Computing: Informatics and Systems.* Doi.org/10.1016/j.suscom.2014.03.002.

[17] M. Dayarathna, Y. Wen, R. Fan. 2016. Data Center Energy Consumption Modeling: A Survey. *IEEE Communications Surveys Tutorials.* 18: 732-794. Doi:10.1109/COMST.2015.2481183

[18] P. L'Ecuyer, L. Meliani, J. Vaucher. 2002. SSJ: A Framework for Stochastic Simulation in Java. E. Yücesan, C.-H. Chen, J. L. Snowdon, J. M. Charnes (Eds.). *Proceedings of the 2002 Winter Simulation Conference*. IEEE Press. 234-242.

[19] C. Reiss, J. Wilkes, J. L. Hellerstein. Nov 2011. Google Cluster-usage Traces: Format + Schema. Technical report, Google Inc., Mountain View, CA, USA.

[20] Nist/sematech e-handbook of statistical methods (April 2012). URL http://www.itl.nist.gov/div898/.

[21] S. Makridakis, S. C. Wheelwright, R. J. Hyndman. 1998. *Forecasting: Methods and Applications.* 3rd Edition. John Wiley & Sons, Inc.

[22] J. E. Vera, S. F. Mora and R. A. Cervantes. 2016. Design and Testing of a Network of Sensors on Land Surfaces to Prevent Landslides. *2016 IEEE Biennial Congress of Argentina (ARGENCON), Buenos Aires, Argentina*. 1-4. Doi: 10.1109/ARGENCON.2016.7585271.

[23] Portal información-transacciones. 2017. URL http://www.xm.com.co/Pages.

[24] Demanda y eficiencia energética. URL http://www1.upme.gov.co.

[25] Power and performance data sheet-dell poweredge r510 featuring the dell energy smart 750w psu and intel xeon x5670 (2009). URL www.dell.com.

[26] C. Mendoza, A. Quintero, F. Santamaria, J. Alarcon. 2016. Estimation of Electric Energy Required by Electric Vehicles based on Travelled Distances in a Residential Zone. *Tecciencia*. 11(21). Doi: http://dx.doi.org/10.18180/tecciencia.2016.21.4.