
# II Jornadas de Doctorado en Informática

14 de febrero de 2019, Universidad de Almería

---

Programa de Doctorado en Informática

Departamento de Informática

Escuela Internacional de Doctorado de la Universidad de Almería

Programa de Doctorado en Informática
Departamento de Informática
Escuela Internacional de Doctorado de la Universidad de Almería
Universidad de Almería

# II Jornadas de Doctorado en Informática

14 de febrero de 2019, Universidad de Almería

---

Las II Jornadas de Doctorado en Informática de la Universidad de Almería (JDI2019@UAL) es un evento que se realiza en el marco del Programa de Doctorado en Informática de la UAL cuyo objetivo principal es la realización del seguimiento de la investigación anual de los estudiantes de doctorado, y que sirve, a la vez, como punto de encuentro anual para estudiantes, tutores, directores y miembros de la comunidad universitaria con intereses en la investigación científica en el ámbito de la Informática. Así mismo, estas jornadas pretenden servir como foco para la difusión y divulgación de los resultados de investigación de la Informática que se están desarrollando en las tesis doctorales soportadas por los actuales proyectos de I+D de los Grupos de Investigación del Departamento de Informática.

En su 2da edición, estas jornadas ha contado con la presentación de los resultados de la tesis doctoral "*A Recommender System for Smart User Interfaces using Machine Learning and Microservices*" presentada por el estudiante de doctorado Antonio Jesús Fernández. Además, esta edición también ha contado con la conferencia invitada titulada "*Analítica Predictiva en la gestión inteligente de redes de distribución de agua potable: un caso real*" impartida por el profesor Juan Carlos Preciado de la Universidad de Extremadura.

https://sites.google.com/ual.es/jdi2019/


Programa de Doctorado en Informática
Departamento de Informática
Escuela Internacional de Doctorado de la Universidad de Almería
Universidad de Almería

II Jornadas de Doctorado en Informática
Edita: Comisión Académica de Doctorado en Informática
Lugar: Universidad de Almería
https://sites.google.com/ual.es/jdi2018/
@doctoradoINFUAL
doctorado.informatica.ual@gmail.com

## SECCIÓN I: Trabajos presentados



## SECCIÓN II: Otros trabajos

# SECCIÓN I

# Trabajos presentados

---

1. Alamin, Yaser: "Artificial Neural Network models to predict energy".
2. Calvo Cruz, Nicolás: "Computación de Altas Prestaciones en el Diseño Óptimo y Control de Plantas Solares de Torre. Control Óptimo del Campo de Helióstatos".
3. Carballo López, José Antonio: "Modelado y Optimización para una Gestión eficiente de Recursos en tecnología termosolar. Modelado y optimización en termosolar".
4. Fernández García, Antonio Jesús: "A Recommender System for Smart User Interfaces using Machine Learning and Microservices".
5. García García, Francisco José: "Tratamiento de Spatial Big Data: Técnicas de Particionado y Procesamiento de Consultas Espaciales".
6. Gil Vergel, Juan Diego: "Aportaciones desde el punto de vista del modelado y del control automático a la tecnología de destilación por membranas alimentadas con energía solar".
7. Moreno Riado, Juan José: "Aceleración del filtro basado en Difusión No-Lineal Anisótropa".
8. Ojeda Castelo, Juan Jesús: "El Modelo Dispositivo - Interacción y Machine Learning en Interacción Natural".
9. Orts Gómez, Francisco José: "Optimizando la eficiencia energética de SMACOF".
10. Puertas Martín, Savíns: "Cribado virtual aplicado al potencial electrostático usando un algoritmo evolutivo".
11. Ramos Teodoro, Jerónimo: "Gestión de recursos heterogéneos en <<energy hubs>> con autoconsumo".
12. Ruiz Ferrandez, Miriam: "Calibrado de parámetros en modelos epidemiológicos complejos: una aproximación multi-objetivo".

---

# Artificial Neural Network models to predict energy

Yaser Imad Alamin

[1] Department of Informatics, University of Almería, Agrifood Campus of International Excellence (ceiA3) CIESOL Joint Centre University of Almería - CIEMAT, 04120 Almería, Spain. (e-mail: ya312@inlumine.ual.es, jhervas@ual.es and mcastilla@ual.es)

**Abstract.** Climate change, the decrease in fossil-based energy resources and the need of reducing the greenhouse gas emissions require energy efficient and smart buildings. Moreover, the ratio of renewable energy sources should be increased against traditional energy sources. Therefore, models are necessary to predict the behaviour in buildings' energy consumption. In this work, three different Artificial Neural Networks (ANNs) models have been developed. Firstly, a model which allows predicting Electrical Load Demand (ELD) for the Heating, Ventilating and Air Conditioning (HVAC) system which is extremely important for its management. Secondly, an ANN to predict the energy generated by a Concentrator PhotoVoltaic (CPV) systems, that is located at University of Rabat (Morocco), has been also developed. More in detail, the previous models can be used as efficient strategies in order to reach a Net Zero Energy Buildings (NZEB). Finally, a model is being developed to assess thermal transmittance in walls in two identical building located at Holzkirchen, south to Munich in Germany, in order to give more attention to the actual energy consumption of new buildings.

**Keywords:** Artificial Neural Networks · HVAC · comfort control · Concentrator PhotoVoltaic · Estimation of thermal transmittance.

## 1 Introduction

Most people spend more than 90% of their time inside buildings, and almost 50% of the energy consumption is used to obtain suitable thermal comfort conditions in commercial buildings. Therefore, the development of HVAC systems that do not rely on fossil fuels with a higher energy-efficient is important to reducing energy consumption [32]. Therefore, the prediction of the Electrical Load Demand (ELD) in the target building or the specific target system in the building

is being widely studied nowadays since optimisation energy and balancing with the use of renewable sources through specific control systems requires accurate knowledge of energy consumption profile in the building [10].

AI methods are a research line that has been experiencing an increasing focus over the past years because of their good fit for this kind of problems. Nowadays AI is used for almost everything, particularly in control systems. Among their uses, it is possible to highlight to get and maintain the users' thermal comfort in the buildings. AI includes several techniques as ANNs, fuzzy logic, support vector machine, genetic programming or a combination of them which is also well-known as hybrid systems [10], [18], [29], [9], [2]. A grey-model can be used to analyse the energetic behaviour of the building when only incomplete or uncertain data are available [26]. Here some of the physical facts are introduced in the system and the other calculated by AI.

To date, the rising cost of fossil fuels electricity production and the challenge of reducing the carbon emissions are favourably shifting energy production to clean energy sources such as solar PhotoVoltaics (PV). Photovoltaics represent the best known solar energy system and more specifically, Concentrator Photo-Voltaic (CPV) is a promising renewable energy technology that is able to reduce the fossil fuel dependence [30]. CPV uses cheap optical devices like curved mirrors and lenses to focus solar direct radiation onto small, but highly efficient multi-junction solar cells. Solar tracker and cooling systems are part of a standard CPV facility [24].

Power production instability can cause operation and control issues for energy suppliers and users. Thus, power forecasting represents a key factor since eventual sudden increases or drops in the power production can be predicted. This may allow the energy supplier to be able to regulate his services by shifting, for example, the PV use to other energy sources [16]. Currently, the short-term forecasts of the output power of PV systems can be done either directly or indirectly [33], [31]. The difference between the two main methods resides in the fact that indirect forecast, there is no need for solar radiation forecasts as an intermediate step where environmental parameters such as ambient temperature, wind speed, wind direction, relative humidity and clearness index should be considered in the calculations [11], [20]. This makes a direct forecast more accurate than an indirect one. Direct PV power forecasting methods can be based on empirical equations or on machine learning techniques. While using machine learning models, the inputs of the irradiance modelling can directly be applied to forecast the power. A complete review of the PV power forecasting can be found in [4].

AI models have been used in PV applications [21] since they can identify the system dynamics without explicitly knowing the interactions between its components. ANNs are currently used for the modelling of solar power and energy systems in a wide range application both in the demand side [2] and the production ones[19]. Feedforward or radial basis function artificial neural networks are involved in typical applications [7].

In addition, ANNs can be combined with other AI techniques such as Genetic Algorithms (GA) [23], Particle Swarm Optimisation (PSO), Genetic Swarm Optimisation (GSO) [22], firefly optimisation [5], stepwise regression [25], fuzzy logic [28] or Principal Component Analysis (PCA) [17] in order to optimise the set of inputs, the complexity in the modelling of the system makes ANNs suitable.

In this work, several ANNs models are presented, firstly a model to predict the consumption of the HVAC system used in CIESOL building have been developed. Prediction models based on ANNs, which have been chosen for their distinctive features for this problem, have been obtained. Among them, two models have been selected for two seasons, summer and winter. Secondly a model has been developed for the prediction of the maximum power of a CPV system, since the ANNs are capable to obtain models for the system without knowing the details of the system, a simple knowledge of the variables and its effect on the output of the system, the ANNs can be presented as black box [34], [13].

Moreover, the third model is being developed to estimate the thermal characterisation of the walls of buildings using the same method, using ANNs [6] and grey-box models [8] to predict thermal behaviour of buildings and walls showing promising results in the literature, and give good predictions of building thermal behaviour, as such would be suitable for model predictive control [14].

## 2 Scope of the research

### 2.1 CIESOL building

In the work presented in [1], a model to predict the consumption of the fan-coil system that is used to maintain the users' thermal comfort conditions inside the rooms of the CIESOL building is presented. The CIESOL building (http://www.ciesol.es) is a solar energy research centre placed on the Campus of the University of Almería in the Southeast of Spain.

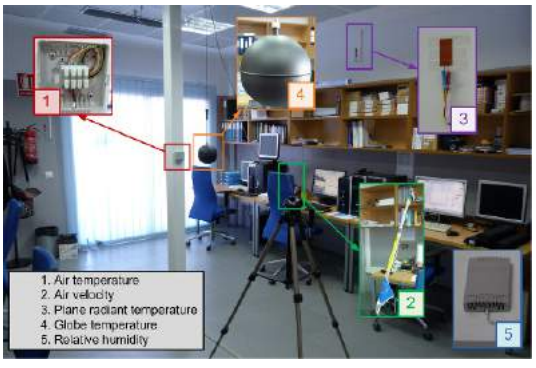



**Fig. 1.** Sensors network in a room of the CIESOL building

More specifically, this building is divided into two different floors with a total surface approximately equal to 1072 $m^2$. In addition, it has been designed to be a Nearly-Zero Energy Building (NZEB) and, thus, it has been designed following several bioclimatic criteria such as the use of photovoltaic panels to produce electricity and an HVAC system based on solar cooling which is composed of a solar collector field, a hot water storage system, a boiler and an absorption machine with its refrigeration tower. Therefore, this HVAC system is able to produce heat or cold air for the whole building as a function of the demanded needs. To do that, hot or cold water flows inside the building and, at each room, it goes through a fan-coil unit. This fan-coil unit allows to introduce air at a certain temperature inside each room by regulating both the amount of water which flows through it (by means of a two-way valve), and the volume of air which is introduced in the room (by means of a three-position fan), see Figure 2.

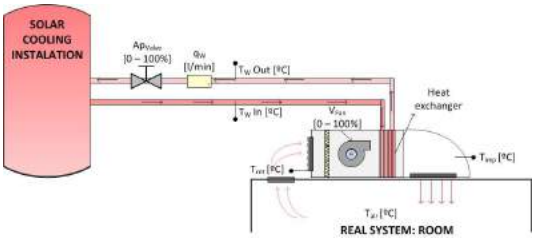



**Fig. 2.** Scheme of a fan-coil unit

## 2.2 Photovoltaic

The second model presented in [3] which is used to predict the output energy of the Hight Concentrator PhotoVoltaic (HCPV). The HCPV facility consists of 108 CPV modules mounted in three strings of 36 modules and connected in series. The CPV facility is located on the campus of International University of Rabat (UIR) in the middle-West of Morocco (geographical coordinates: latitude 33.982 N, longitude 6.7248 W). Figure 3 shows the HCPV plant and the technical characteristics of the modules provided by the manufacturer are displayed in Table 1.

An SHP1 Kipp and Zonen pyrheliometer have been settled on the solar tracker in order to measure the DNI. And weather station located nearby was used to record other environmental parameters, such as wind speed, air temperature, relative humidity and wind direction. While the Solar elevation (h) is deduced using a dedicated software installed on the tracker.

**Fig. 3.** The HCPV facility under study

| Module specifications | |
|---|---|
| Primary Optics | PMMA Fresnel lens |
| Dimension of Primary Optics | 31cm x 31cm |
| Secondary optics | Refractive truncated pyramid |
| Type of solar cells | Lattice-matched GaInP/GaInAs/Ge |
| Cell dimension | 1cm x 1cm |
| Geometrical concentration | X 961 |
| Concentration | X 800 |
| Number of solar cells | Six cells in series |
| Module maximum power | 110 W |
| Open-circuit voltage | 17.70 V |
| Short-circuit current (Isc) | 8,65A |
| Cooling system | Passive |
| Type of cell protection | Bypass diode |

**Table 1.** Characteristics of the HCPV modules used

### 2.3 Identical buildings

The third model is being now developed to predict the thermal transfer of a wall of Twin (Identical) building in Holzkirchen, Germany. The buildings were checked with one another in a sidebyside test and shown to have almost identical performance in terms of heating required to maintain a set temperature and in air leakage etc. Figure 4 shows the building and weather station.

- Heating of the two buildings to constant room temperatures (cellar: 19.5 C, ground floor: 25 C, attic: no heating)
- building services deactivated (gas boiler, ventilation)
- Heating with electrical radiators
- No internal heat sources
- Roller blinds closed (higher heating power without of solar gains)


**Fig. 4.** Twin building and the weather station locations

## 3 Architecture and Methodology

### 3.1 Artificial Neural Networks

ANNs mimic the human brain's biological neural network in the problem-solving process. An ANNs can be seen as a black-box that connects the input to the output, with fully connected neurons (nodes), these nodes being connected by weights. They are used for the non-linear mapping between the input data, X, and the output vector, Y, in order to model relations or to detect patterns among them. Less knowable of the system is required in this case. Hence, ANNs could be very useful in complicated systems and models. By using supervised training methods, the parameters (weights and biases) and structure can be determined from data, Figure 5 shows RBF ANNs scheme.

**Radial Basis Functions** One of the less complicated ANNs is the radial basis functions (RBF), The RBF ANNs is built up of three layers: the input, hidden and output layers respectively. Each layer is fully connected to the previous one by means of nodes. In the input layer, a node is assigned to each of the input variables. The input signal passes directly to the hidden layer without weights. The hidden units contain the RBF, also called transfer functions. They are similar to the sigmoid functions commonly used as activation functions in the back-propagation network models.

The RBF is analogous to the Gaussian density function which is defined by a centre position and a radius or width parameter. The highest output is given by the Gaussian function when the input variables are closest to the centre position. On the other hand, the function decreases monotonically with the distance increase from the centre. The radius defines whether the RBF function should decrease quickly or slowly. The decrease will be quick when the radius is small but slow for large width value. The Gaussian function which is well known as a radially-symmetric function to activate the n neurons of the hidden layer, it can be defined as:

$$f_i(X_r) = exp\left(-\frac{||C_i - X_r||^2}{2\sigma_i^2}\right) \quad , \qquad i = 1, 2, 3...n \tag{1}$$

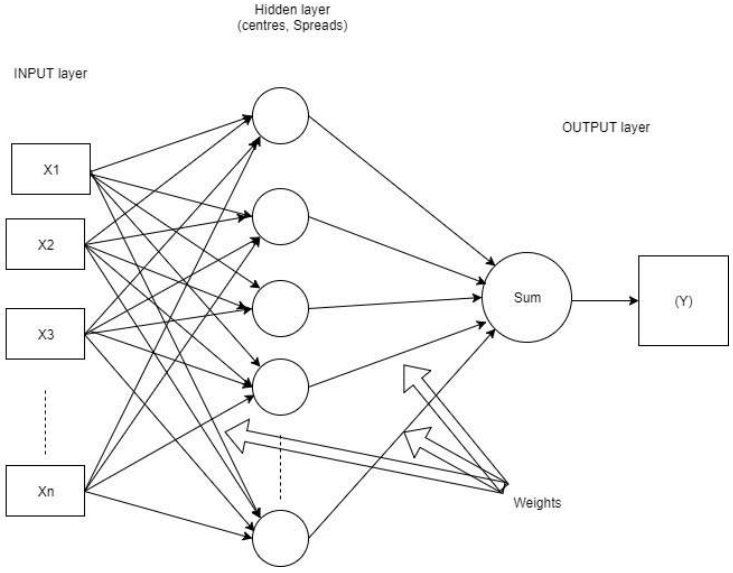



**Fig. 5.** RBF ANNs scheme

In expression 1, $C_i$ is the centre of $ith$ RBF unit, $X_r$ is the input vector, $\sigma_i$ is the width (radius) and n is the number of nodes in the hidden layer. The output value $Y_k$ of an RBF ANNs is the summation of the weighted outputs of the hidden units and the biasing term of the output node. It and can be expressed as:

$$Y_k(X_r) = \sum_{i=1}^{n} W_{ik} f_i(X_r) \quad , \qquad k = 1, 2, 3...m \tag{2}$$

where $w_{ik}$ is the weight corresponding to the connection between the $i^{th}$ RBF unit to the $k^{th}$ output.

RBF ANNs are able to solve nonlinear problems since unsupervised learning in the hidden layer is combined with supervised learning in the following layer. This makes them suitable for photovoltaic power forecasting.

### 3.2 Proposed prediction models for short term power forecasting

RBF ANNs training is performed using a gradient-based algorithm, which minimises the training error. The training process will be terminated when the minimum of the generalisation error is obtained, the error obtained in an unseen data (not training data), as training evolves, a spread data-set used for this process called generalisation data-set. This scheme is a way to solve the problem known as over-training. To compare different trained models, with possible different model structure, a third data-set, denoted as a testing data-set, is needed.

Hence, three different sets of data are used: i) training, ii) generalisation and, iii) testing data-sets [12, 15]. The Root Mean Square Error (RMSE) had been used during the training, while an RMSE per RMS of the power has been used to present during the work.

## 4 Data and Experiments

The data construction for CIESOL building is as follow: A set of historical data from the CIESOL building has been collected. Specifically, the historic data-set comprises of one year since the $1^{st}$ of April 2013 to $31^{st}$ of March 2014. It is composed of 365 days, with a sample time of 1 minute. To obtain a more accurate model, all weekend samples are removed from the data-set, since there is nobody in the room, implying that the HVAC will be off all the time. Besides that, the original data-set is split into two data-sets: one for the summer season and another for the winter one, which will have different properties. For summer, the months of June, July, and September have been considered, August is a holiday period in Spain, thus, it has not been taken into account. In a similar way, for winter, the months of December, January, and February have been considered.

Two delayed values of the output (the fan-coil energy) are used as ANNs inputs. The exogenous inputs are the impulse air velocity, delayed by one sample and the current value of the inside air temperature.

Due to the sample time of the historical data, the power consumption signal has a random white noise. Therefore, to remove this noise a smooth filter has been used. After that, both data-sets are divided into a training subset with 36% of the total samples, a generalisation subset of 24%, and a testing subset of 40%.

It is necessary to determine which of the obtained ANNs is the best ones. For this aim, the Normalised Root Mean Square Error (NRMSE) has been used. This index is the percentage of the Root Mean Square Error (RMSE), for the prediction horizon of one step. The experiments had been run for each data subsets separately, for all the combinations of these parameters: i) the number of the centres can be 3, 6, 9, 12 and 15 and, ii) the $\lambda$ parameter can be 0.05, 0.01, 0.005 and 0.001. This parameter is the termination criterion with early stopping since an early stopping method with generalisation data-set has been used, it is normally interrupted the training when the resolution parameter specified is achieved before the iteration met, for better understanding and more details see [27].

The data available for HCPV system consists of 92 days from 2016 to 2018. The data was divided into two part of 46 days each (sunny days and cloudy days). Partially cloudy days have been considering as cloudy days, after filtering the data from noise and errors (measurement errors). Three data-sets have been created for each part, training, generalisation and testing. In this work, the inputs used were environmental parameters namely: weather: $T_{air}$ (Air Temperature), $W_S$ (Wind speed), $W_d$ (Wind direction), $A_m$ (Air mass) based on Solar Elevation (measured or calculated), Azimuth angle of the tracker, DNI and the output power of the HCPV system delayed one time step and two time steps. While

in the hidden layer, a wide number of nodes have been tested (from 10 to 20 nodes), and finally with one output node which is the output power of the HCPV system.

The sampling time of the data is 1 minute; thus the taring has been carried out with 1 minute, and another training has been carried out after reforming the data with 5 minute sampling time. The data divide into data-sets, in the same way, did for the data of the fain-coil system mentioned above.

While the data available for training the heat transfer for the walls of the twin building is from one of the twin buildings, and it has taken at two times of the year, mid-August to the end of the September, and from the second weak of April to the end of May. In total, it is 82 days, with time sample of 10 minutes. The data collocated include Air Temperature, Ambient Temperature, Solar Radiation from different directions vertically.

## 5 Results and discussion

From the training of the ANN with real data from the bioclimatic building, different models have been obtained. Table 2 shows some statistics for the best five obtained models for summer for HVAC system using, besides that the NRMSE index, these indicators: Mean Absolute Error (MAE), Mean Relative Error (MRE), Maximum Absolute Error (MaxAE), Standard Deviation Error (StDE) and Normalized Mean Absolute Error (NMAE) for summer. Due to the lack of space only results from the summer case are showed. As it has been pointed out before, the NRMSE index has been used to assess the performance on the different models. Moreover, validation results for one step ahead (using 1-minute interval) show an appropriate performance with an NRMSE less than 0.99% for the worst case. Specifically, the best model for summer is the first one as it is the less complex NN structure with the best accuracy. On the other hand, for the winter case, the best model has 0.6% in the worst case. As in the previous case, the best model is the first one, as well it has the smallest complexity achieving the best accuracy.

Figure 6 shows the prediction of power demand for the fan-coil for the best model, the model 1, for 1, 5, 10 and 15 steps ahead. Due to the lack of space only results from the summer case are showed. As the reader can see in figure the performance of the models decrease when the number of steps ahead predictions increases.

As for HCPV facility model two type of training had been carry on, first taking into account 1-minute sampling time, the best models of the sunny days and best models of the cloudy days are shown in table 3 and table 4, respectively, selected from all the models were trained by using RMSE and choosing the least RMSE to RMS of the power (RMSP) ratio for one step ahead. For the sake of space only results from the 1-minutes sampling time case can be showed.

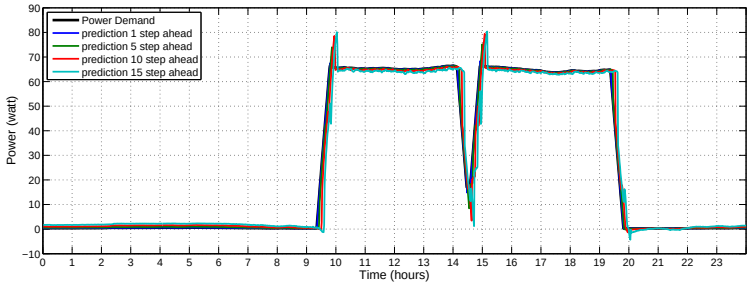



**Fig. 6.** Model 1, 15 steps ahead power demand prediction for summer

| No. | steps ahead | NRMSE | MAE | MRE | MaxAE | StDE | NMAE |
|---|---|---|---|---|---|---|---|
| 1 | 1st | 0.0099 | 0.0579 | 0.3450 | 7.0046 | 0.2209 | 0.0073 |
| | 15th | 0.2975 | 3.7176 | 46.8871 | 86.0279 | 6.4472 | 0.4687 |
| 2 | 1st | 0.0096 | 0.0450 | 0.2199 | 7.2965 | 0.2151 | 0.0057 |
| | 15th | 0.2098 | 2.1173 | 16.3099 | 80.5860 | 4.6304 | 0.2669 |
| 3 | 1st | 0.0097 | 0.0471 | 0.1707 | 7.2523 | 0.2166 | 0.0059 |
| | 15th | 0.2607 | 2.9095 | 23.9566 | 83.4681 | 5.7338 | 0.3668 |
| 4 | 1st | 0.0098 | 0.0522 | 0.3141 | 7.3882 | 0.2191 | 0.0066 |
| | 15th | 0.2359 | 3.0900 | 40.1141 | 85.7121 | 5.2633 | 0.3895 |
| 5 | 1st | 0.0097 | 0.0464 | 0.2462 | 7.4104 | 0.2172 | 0.0058 |
| | 15th | 0.2058 | 1.5047 | 10.2133 | 88.6098 | 4.5983 | 0.1897 |

**Table 2.** Statistical analysis of the best five obtained models for summer

| Model | RMSE | RMSE/RMSP | RMSE 16 steps ahead | RMSE16/RMSP |
|---|---|---|---|---|
| 1 | 573.3 | 0.1607 | 1477.5 | 0.1934 |
| 2 | 572.9 | 0.2483 | 3238.8 | 0.4239 |
| 3 | 573.6 | 0.1486 | 1379 | 0.1805 |
| 4 | 573.7 | 0.1481 | 1423.4 | 0.1863 |
| 5 | 574.1 | 0.1902 | 1802.5 | 0.2359 |

**Table 3.** 1-minute sampling time sunny days models

| Model | RMSE | RMSE/RMSP | RMSE 16 steps ahead | RMSE16/RMSP |
|---|---|---|---|---|
| 1 | 1047.7 | 0.1615 | 1752600 | 27.0236 |
| 2 | 1045.7 | 0.1612 | 2493.3 | 0.3845 |
| 3 | 1046.4 | 0.1613 | 2429.4 | 0.3746 |
| 4 | 1045 | 0.1611 | 2426.3 | 0.3741 |
| 5 | 1041.7 | 0.1606 | 2850.8 | 0.4396 |

**Table 4.** 1-minute sampling time cloudy days models

## 6 Conclusions and Future Work

The research lines developed for the PhD candidate during the last year deals with the development of several NN models to predict energy consumption or producing. For the case of the energy demand of the CIESOL building obtained RBF ANN show optimistic results with a simple structure as the best method for the prediction of the electric load for an HVCA system. Moreover, the prediction of the electric producing by the CPV facility has shown equally optimistic results with the same method. The fact that the developed RBF ANN model is very simple and the computational resources for its application are tiny and easily available at modern automation systems, gives to the model its advantage to be used in several fields. In particular, in order to apply it to a control system, only data from simple sensors and electric power measurements are required.

A future research line is the use of the ANN as the basis of a control system which, through the ANN model, will be able to maintain the thermal comfort of the users of building whereas the energy consumption necessary to reach this thermal comfort situation is minimised.

Following this research line, another ANN will be developed in order to predict the power lost in the wells of a building, this will make the construction of the building more efficient, moreover, Will give a better view to HVCA control for the buildings as well for the PV systems will be needed.

## References

1. Alamin, Y.I., Álvarez, J.D., del Mar Castilla, M., Ruano, A.: An artificial neural network (ANN) model to predict the electric load profile for an HVAC system. IFAC-PapersOnLine **51**(10), 26–31 (2018)
2. Alamin, Y.I., Castilla, M.d.M., Álvarez, J.D., Ruano, A.: An economic model-based predictive control to manage the users thermal comfort in a building. Energies **10**(3), 321 (2017)
3. Anaty, M.K., Alamin, Y.I., Bouziane, K., García, M.P., Yaagoubi, R., Hervás, J.D.Á., Belkasmi, M., Aggour, M.: Output power estimation of high concentrator photovoltaic using radial basis function neural network. International Renewable and Sustainable energy Conference (2018)
4. Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F., Antonanzas-Torres, F.: Review of photovoltaic power forecasting. Solar Energy **136**, 78–111 (2016)
5. Ashraf U. Haque, M. Hashem Nehrir, P.M.: Solar pv power generation forecast using a hybrid intelligent approach. IEEE Power and Energy Society General Meeting (PES) pp. 1–5 (2013). https://doi.org/10.1109/pesmg.2013.6672634
6. Bienvenido-Huertas, D., Moyano, J., Rodríguez-Jiménez, C.E., Marín, D.: Applying an artificial neural network to assess thermal transmittance in walls by means of the thermometric method. Applied Energy **233**, 1–14 (2019)
7. Bonanno, F., Capizzi, G., Graditi, G., Napoli, C., Tina, G.M.: A radial basis function neural network based approach for the electrical characteristics estimation of a photovoltaic module. Applied Energy **97**, 956–961 (2012)
8. Brastein, O., Perera, D., Pfeifer, C., Skeie, N.O.: Parameter estimation for grey-box models of building thermal behaviour. Energy and Buildings **169**, 58–68 (2018)

9. Calvino, F., La Gennusa, M., Rizzo, G., Scaccianoce, G.: The control of indoor thermal comfort conditions: introducing a fuzzy adaptive controller. Energy and buildings **36**(2), 97–102 (2004)
10. Castilla, M.d.M., Álvarez, J.D., Rodríguez, F., Berenguel, M.: Comfort control in buildings (2014)
11. Chow, S.K., Lee, E.W., Li, D.H.: Short-term prediction of photovoltaic energy generation by intelligent approach. Energy and Buildings **55**, 660–667 (2012)
12. Ferreira, P.M., Ruano, A.E.: Exploiting the separability of linear and nonlinear parameters in radial basis function networks. In: Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000. pp. 321–326. IEEE (2000)
13. Hadjiiski, L., Geladi, P., Hopke, P.: A comparison of modeling nonlinear systems with artificial neural networks and partial least squares. Chemometrics and intelligent laboratory systems **49**(1), 91–103 (1999)
14. Harish, V., Kumar, A.: A review on modeling and simulation of building energy systems. Renewable and Sustainable Energy Reviews **56**, 1272–1292 (2016)
15. Haykin, S.: Neural Networks A comprehensive Foundation. PEARSON Prentice Hall, Ontario, Canada (2005)
16. Hernandez, L., Baladron, C., Aguiar, J.M., Carro, B., Sanchez-Esguevillas, A.J., Lloret, J., Massana, J.: A survey on electric power demand forecasting: future trends in smart grids, microgrids and smart buildings. IEEE Communications Surveys & Tutorials **16**(3), 1460–1495 (2014)
17. Junior, J.G.d.S.F., Oozeki, T., Ohtake, H., Shimose, K.i., Takashima, T., Ogimoto, K.: Regional forecasts and smoothing effect of photovoltaic power generation in japan: an approach with principal component analysis. Renewable Energy **68**, 403–413 (2014)
18. Lopez, A., Sanchez, L., Doctor, F., Hagras, H., Callaghan, V.: An evolutionary algorithm for the off-line data driven generation of fuzzy controllers for intelligent buildings. In: Systems, man and cybernetics, 2004 IEEE international conference on. vol. 1, pp. 42–47. IEEE (2004)
19. Mandal, P., Madhira, S.T.S., Meng, J., Pineda, R.L., et al.: Forecasting power output of solar photovoltaic system using wavelet transform and artificial intelligence techniques. Procedia Computer Science **12**, 332–337 (2012)
20. McVey-White, P., Besson, P., Baudrit, M., Schriemer, H.P., Hinzer, K.: Effects of lens temperature on irradiance profile and chromatic aberration for cpv optics. In: AIP Conference Proceedings. p. 040004. No. 1, AIP Publishing (2016)
21. Mellit, A., Kalogirou, S.A., Hontoria, L., Shaari, S.: Artificial intelligence techniques for sizing photovoltaic systems: A review. Renewable and Sustainable Energy Reviews **13**(2), 406–419 (2009)
22. Ogliari, E., Grimaccia, F., Leva, S., Mussetta, M.: Hybrid predictive models for accurate forecasting in pv systems. Energies **6**(4), 1918–1929 (2013)
23. Pedro, H.T., Coimbra, C.F.: Assessment of forecasting techniques for solar power production with no exogenous inputs. Solar Energy **86**(7), 2017–2028 (2012)
24. Pérez-Higueras, P., Fernández, E.F.: High concentrator photovoltaics: fundamentals, engineering and power plants. Springer (2015)
25. Ramsami, P., Oree, V.: A hybrid method for forecasting the energy output of photovoltaic systems. Energy Conversion and Management **95**, 406–413 (2015)
26. Rodríguez, F., Berenguel, M., Arahal, M.: A hierarchical control system for maximizing profit in greenhouse crop production. In: European Control Conference (ECC), 2003. pp. 2753–2758. IEEE (2003)

27. Ruano, A., Ferreira, P., Fonseca, C.: An overview of nonlinear identification and control with neural networks. IEE Control Engineering Series **70**, 37 (2005)
28. Simonov, M., Mussetta, M., Grimaccia, F., Leva, S., Zich, R.: Artificial intelligence forecast of pv plant production for integration in smart energy systems. International Review of Electrical Engineering **7**, 3454–3460 (02/2012 2012)
29. Singh, J., Singh, N., Sharma, J.: Fuzzy modeling and control of hvac systems–a review (2006)
30. Swanson, R.M.: The promise of concentrators. Progress in Photovoltaics: Research and Applications **8**(1), 93–111 (2000)
31. Tao, C., Shanxu, D., Changsong, C.: Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement pp. 773–777 (2010)
32. Vakiloroaya, V., Samali, B., Fakhar, A., Pishghadam, K.: A review of different strategies for hvac energy saving. Energy Conversion and Management **77**, 738–754 (2014)
33. Yang, H., Huang, C., Huang, Y., Pai, Y.: A weather-based hybrid method for 1-day ahead hourly forecasting of pv power output. IEEE Transactions on Sustainable Energy **5**(3), 917–926 (July 2014). https://doi.org/10.1109/TSTE.2014.2313600
34. Zhang, G.P., Patuwo, B.E., Hu, M.Y.: A simulation study of artificial neural networks for nonlinear time-series forecasting. Computers & Operations Research **28**(4), 381–396 (2001)

# Computación de Altas Prestaciones en el Diseño Óptimo y Control de Plantas Solares de Torre

## Control Óptimo del Campo de Helióstatos

N.C. Cruz[1]

Universidad de Almería, ncalvocruz@ual.es

**Resumen** El campo de helióstatos de las plantas solares de recepción central debe configurarse con precisión para i) maximizar el aprovechamiento de la radiación solar incidente y ii) mantener el sistema en un estado de operación seguro. Las tareas de control implican tanto decidir qué helióstatos del campo activar como hacia dónde apuntarlos sobre el receptor. Esta tarea entraña una gran complejidad porque las plantas actuales cuentan con cientos, incluso miles, de helióstatos. En este contexto, es importante lograr ciertas distribuciones de flujo deseables sobre el receptor para garantizar un funcionamiento óptimo y evitar problemas de estrés térmico y daños. En este trabajo se presentan una estrategia para modelar analíticamente el comportamiento del campo y una meta-heurística capaz de usar dicho modelo para reproducir cualquier distribución de flujo sobre el receptor. Esta última es capaz de seleccionar y apuntar el conjunto óptimo de helióstatos a usar minimizando una función de error. Este documento informa de los avances en un proyecto de tesis doctoral y obvia ciertos detalles bajo gestión de varias revistas científicas.

## 1. Introducción

Las centrales solares de recepción central (CSRC) son uno de los tipos de instalación más interesante y prometedor para la generación de electricidad a gran escala. Esto es debido a su eficiencia termodinámica, la madurez de su base tecnológica y una cierta estabilidad de producción [5,11].

Una central CSRC se compone, a grandes rasgos, de un conjunto de espejos altamente reflectantes llamados helióstatos y de un receptor de radiación ubicado normalmente a gran altura en una torre. Los espejos tienen una estructura orientable y siguen la trayectoria aparente del Sol a lo largo del día para concentrar la radiación incidente sobre el receptor. Éste alcanza una gran temperatura que es transferida progresivamente a un fluido caloportador que circula en su interior. Cuando la temperatura de dicho fluido es suficientemente elevada puede usarse finalmente en un ciclo de turbina clásico. La figura 1 muestra un esquema sencillo de una CSRC. El lector interesado puede consultar los trabajos de [1,2,10] para obtener más información acerca de esta tecnología.

El campo de helióstatos de las centrales CSRC cuenta con muchas unidades desplegadas, especialmente considerando que se suele sobredimensionar para

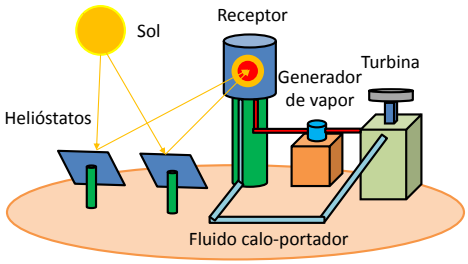



**Figura 1.** Esquema de una planta solar de recepción central.

hacer frente a condiciones desfavorables como días nublados. Sin embargo, no tienen por qué ser necesarios siempre todos los helióstatos para alcanzar la potencia nominal. De hecho, no se debe exponer el receptor a una densidad de radiación excesiva o descontrolada sobre su superficie. La distribución de flujo que el campo de helióstatos proyecta sobre el receptor debe controlarse cuidadosamente. Es necesario evitar gradientes de temperatura muy pronunciados sobre el receptor que puedan causar estrés térmico y el envejecimiento prematuro de sus componentes [4,3,8,12]. Éste es un factor clave para aumentar la vida útil del receptor, lo que influye directamente en los costes de producción de las centrales CSRC como se destaca en [8].

La distribución o mapa de flujo que se forma sobre el receptor depende directamente de los helióstatos activos y sus puntos de enfoque. En este contexto, para configurar una planta CSRC se puede plantear un problema de optimización de dos capas en el que hay que decidir i) qué helióstatos activar de entre todos los disponibles y ii) a qué punto del receptor se enfoca cada uno de ellos. Estas tareas están normalmente basadas en decisiones manuales de operarios, lo que limita implícitamente el número de puntos de enfoque que pueden gestionar y la adaptabilidad del campo.

En los recientes trabajos de [3,8] se abordan dos problemas de optimización similares y basados en conjuntos predefinidos de puntos de enfoque. Se centran en minimizar la desviación típica de la distribución de flujo y la diferencia entre su máximo y mínimo respectivamente. En [3] se usa con éxito un algoritmo genético mientras que en [8] se logran buenos resultados mediante una búsqueda TABÚ. Sin embargo, este trabajo pretende definir y abordar un caso más general en el cual, dada una forma de flujo deseada en un cierto instante, se determina la configuración completa del campo (helióstatos activos y sus puntos de enfoque) para obtenerla. En lugar de estar ligado a objetivos fijos como homogeneizar el mapa de flujo, se minimiza directamente la diferencia entre el mapa deseado y el que se obtiene mediante un modelo del campo. Además, la asignación de puntos de enfoque se lleva a un espacio continuo. De esta forma, el campo de helióstatos se hace significativamente más configurable. El método desarrollado tiene dos componentes: i) un algoritmo genético para selección de helióstatos y apunte inicial y ii) un descenso de gradiente para ajustar con mayor precisión los puntos

de enfoque una vez se ha fijado el conjunto de helióstatos a activar. Además, dada la necesidad de contar con un modelo analítico del campo a controlar, se ha desarrollado una metodología para crear este tipo de modelos a partir de un conjunto reducido de datos precisos.

## 2. Avances

### 2.1. Planteamiento del problema

Se busca replicar cualquier distribución de flujo deseada sobre un receptor plano seleccionando, de su campo de helióstatos asociado, cuáles se deben activar y hacia dónde deben apuntar. Para tal fin, se plantea un problema de optimización de gran escala, es decir, con un gran número de variables.

El campo de helióstatos puede definirse como un conjunto ordenado de cardinalidad $N$, $H = \{h_1, h_2, \cdots, h_N\}$. El mapa de flujo a obtener en un cierto instante $t$, también conocido como mapa de referencia, se define por una función bidimensional $F$ que determina la densidad de radiación en $kW/m^2$ en cualquier punto $(x, y)$ del receptor. Éste está orientado al norte, y sus direcciones $X$ e $Y$ apuntan hacia el Este y el cielo, respectivamente. Cada helióstato $h$ proyecta una cierta distribución de flujo $f_h$ sobre el receptor cuando está operativo. $f_h$ es también una función bidimensional de densidad de radiación.

Una configuración del campo o solución candidata, $C$, se define como una secuencia de longitud $N$ con la estructura $C = \{c_1, c_2, \cdots, c_N\}$. En $C$, la posición de cada elemento se vincula directamente con un cierto helióstato en $H$. Es decir, $c_h$ define la configuración particular del helióstato $h$, y ésta puede ser $\emptyset$ cuando no está activo o las coordenadas $(x, y)$ de su punto de enfoque sobre el receptor. Por consiguiente, hay $2N^*$ variables a optimizar en la segunda capa del problema, donde $N^*$ es el número de helióstatos activos finalmente, y $N^* \leq N$.

En este contexto, una cierta configuración del campo define la distribución de flujo $F^*$ que se forma sobre el receptor. Ésta se obtiene superponiendo el mapa de flujo asociado a cada helióstato, $f_h$, y descartando aquellos inactivos. Entonces, la función objetivo del problema abordado puede definirse como la diferencia entre la forma deseada o referencia, y el mapa obtnido con una configuración $C$, $O = (F - F^*(C))^2$. Por consiguiente, el problema optimización se define desde una perspectiva de minimización como se muestra a continuación:

$$minO = min\,(F - F^*(C))^2 \tag{1}$$

Asumiendo que las funciones de las distribuciones de flujo son continuas, la expresión 1 implica una discretización de facto de los mapas de flujo de referencia y obtenido. Éstos puede verse como imágenes en escala de grises para ser comparados. Por tanto, después de definir una malla de discretización sobre el plano del receptor, el problema puede formularse de la siguiente forma extendida:

$$minO = min \sum_{x \in X_T} \sum_{y \in Y_T} (F(x, y) - F^*(C)(x, y)) \tag{2}$$

$X_T$ e $Y_T$ son los equivalentes discretos de los ejes asociados a las direcciones $X$ e $Y$ sobre el receptor, respectivamente. Nótese que ambos conjuntos pueden tener distinta cardinalidad, es decir, el plano del receptor no tiene por qué ser cuadrado.

En relación al mapa de flujo que cada helióstato $h$ proyecta sobre el receptor, $f_h$, éste se define analíticamente por una función Gaussiana bidimensional con la siguiente formulación:

$$f_h(x,y) = \frac{P}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right)} \quad (3)$$

$x$ e $y$ son las coordenadas escogidas sobre el plano del receptor al discretizarlo. $P$ es la contribución de potencia del helióstato $h$ sobre el receptor. $\rho$ es la correlación entre la forma del mapa de flujo a lo largo de $X$ e $Y$. $\sigma_x$ y $\sigma_y$ son las desviaciones típicas de los valores de densidad de radiación en las dimensiones $X$ e $Y$, respectivamente. $\mu_x$ y $\mu_y$, que hacen referencia a los valores medios de la distribución Gaussiana, definen el punto central del mapa de flujo, es decir, el punto de enfoque del helióstato $h$. Esta estrategia es similar a la seguida en [3,8], donde se usa una función Gaussiana de base circular siguiendo el modelo HFLCAL [9].

Como se mencionó anteriormente, es necesario saber la forma que cada helióstato proyecta en el receptor cuando se activa. En general, esta información puede obtenerse mediante trazado de rayos, computacionalmente costoso pero muy flexible y preciso, o con modelos analíticos de convolución, menos costosos pero también menos precisos [3,8]. Sin embargo, en este trabajo se opta por generar y almacenar los parámetros $P$, $\rho$, $\sigma_x$ y $\sigma_y$ que definen la forma del mapa de flujo de cada helióstato en el instante de interés, $t$. En la sección 2.2 se resume la metodología diseñada para tal fin.

Finalmente, es importante destacar que el mapa de flujo a obtener, $F$, independientemente de si se definió con una función analítica como la propia expresión 3 o cualquier otra, es visto en última instancia como una imagen o matriz de puntos. Por tanto, la resolución del problema es realmente independiente del conjunto de valores en el que se discretiza $F$.

### 2.2. Modelado del comportamiento de los helióstatos

Se ha desarrollado una metodología general que formaliza el proceso de construir un modelo analítico, $M$, de un campo de helióstatos objetivo. Dicho modelo es capaz de predecir los parámetros de forma (p. ej. $P$, $\rho$, $\sigma_x$ y $\sigma_y$) de cualquier helióstato del campo caracterizado, para cualquier instante (posición solar aparente) del año. La metodología se basa directamente en procesos de análisis y modelado de datos para generalizar el comportamiento de un conjunto de helióstatos en una serie de instantes estudiados. Define cinco pasos consecutivos que se describen a continuación. La figura 2 incluye además un resumen gráfico de la misma.

En primer lugar, se selecciona un subconjunto de helióstatos que cubra todas las zonas del campo. Es importante distribuirlos en zonas límítrofes e intermedias. De cada uno de esos helióstatos seleccionados se registran sus coordenadas dentro del campo así como los mapas de flujo que proyectan sobre el receptor. En principio, éstos se obtienen de forma exclusiva y apuntando al centro del mismo para distintas posiciones solares del año. Dichas posiciones han de escogerse del espacio de movimiento del Sol para la ubicación del campo siguiendo un criterio similar al de la selección de los helióstatos: tener representación de todas las zonas. Se destaca además que los mapas de flujo de cada helióstato pueden obtenerse bien de mediciones reales o bien de simulación precisa.

En segundo lugar, se escoge una expresión analítica con la que se pueda describir correctamente la forma de flujo que proyecta un helióstato. Una buena elección es la expresión 3, que es la que se ha usado en el seno de este trabajo. Tomada dicha decisión a conveniencia, se procede a ajustar cada uno de los mapas de flujo previos a la expresión analítica en cuestión. Es necesario registrar la parametrización resultante para cada helióstato en cada instante, es decir, las distintas tuplas de valores $P$, $\rho$, $\sigma_x$ y $\sigma_y$ en este caso.

En tercer lugar, se debe emparejar el conjunto de parámetros que describen analíticamente cada mapa de flujo con la posición de los helióstatos y la posición solar correspondiente. Se reserva aleatoriamente un conjunto de registros para el paso siguiente y el trabajo en éste se centra en la parte restante. El objetivo de este paso es identificar las relaciones subyacentes entre las variables que identifican las posiciones del sol y los helióstatos con los parámetros que describen los mapas de flujo resultantes. Una vez identificadas, se deben modelar analíticamente de forma que sea posible estimar cualquier parámetro $P$, $\rho$... conocidas las coordenadas de cualquier helióstato del campo y una cierta posicuón solar. El resultado sería un modelo $M$. Éste podría estimar la representación analítica que generaría un helióstato del campo en cualquier instante del año. En este punto son útiles estrategias de ajuste polinómico y aprendizaje automático en general.

En cuarto lugar, se debe evaluar la calidad de $M$, tanto estructural como paramétrica (coeficientes ajustados). Para tal fin, se debe usar $M$ para predecir los mismos valores que se reservaron en el paso anterior y compararlos. Si los parámetros que estima $M$ son suficientemente similares a aquellos obtenidos directamente ajustando los mapas de flujo tras el primer y segundo paso, se considera que el modelo funciona correctamente. Si no se lograra realizar con éxito, habría que revisar los pasos previos.

En quinto y último lugar, $M$ está listo para desplegarse, es decir, para usarse en cualquier aplicación. Puede ser útil para compartir una representación compacta de un campo de interés, tareas de control, generar conjuntos de pruebas personalizados... En el marco de este trabajo, se usaría estimando los parámetros de los mapas de flujo de los helióstatos del campo para la posición solar en la que se quiere obtener una configuración determinada.

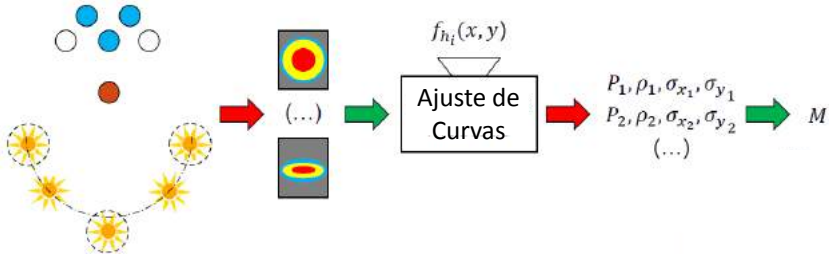



**Figura 2.** Resumen gráfico del proceso de modelado anual de un campo de helióstatos.

La metodología aquí descrita es un resumen de la publicación disponible en [7], realizada en el seno del programa de doctorado, y donde el lector interesado puede encontrar más información.

### 2.3. Método de optimización de dos capas

Para resolver el problema planteado en la sección 2.1 se ha diseñado un método formado por dos capas interconectadas, un algoritmo genético en la primera y un método de descenso por gradiente en la segunda. El algoritmo genético se centra en encontrar el mejor subconjunto de helióstatos a activar para replicar la forma dada. El método de descenso por gradiente parte de la solución inicial del anterior y ajusta los puntos de enfoque dados a aquellos helióstatos fijados como activos. En todo momento se intenta minimizar la diferencia entre el mapa de flujo de referencia y el que se obtiene según el modelo analítico del campo.

En la primera capa, el algoritmo genético genera una población de diferentes soluciones candidatas (vectores $C$) y trabaja sobre ellas haciéndolas evolucionar un cierto número de ciclos dado. De entre las posibles soluciones iniciales, hay tres que siempre se consideran y no son aleatorias: una con todos los helióstatos activos, una con los más potentes para alcanzar el total de la referencia (independientemente de su forma) y otra, con la misma intención que la anterior, pero con los menos potentes. De las dos últimas se extraen cotas para limitar superior e inferiormente el número de helióstatos a admitir en las futuras soluciones candidatas. Tras el proceso de generación inicial se aplica un clásico bucle evolutivo en el que se van creando nuevas soluciones cada ciclo mediante la combinación de las mejores existentes (reproducción) y la inclusión de cambios aleatorios (mutación). Se mantiene una priorización a la supervivencia de aquellas soluciones que dan lugar a una forma más similar a la referencia, es decir, minimizan la función $O$ (presión selectiva).

Si bien el algoritmo genético se centra en escoger elementos (helióstatos), sigue siendo necesario construir un vector de configuración $C$ válido a partir de cada selección candidata. Si no, no es posible valorar el potencial de una cierta selección. Para tal fin, los helióstatos se enfocan iterativamente al pico máximo

de diferencia de potencia entre el mapa de flujo de referencia y el obtenido, empezando por el más potente. Esta estrategia heurística de enfoque dará un criterio de comparación de soluciones con la función objetivo y permitirá obtener un punto de partida válido en el espacio de búsqueda para la segunda capa.

Finalmente, el método de la segunda capa se centra en los helióstatos que se fijaron en la mejor solución del algoritmo genético, y en el punto de enfoque inicial asignado a cada uno. Concretamente, calcula el vector gradiente de la función objetivo, que indica la dirección de máximo aumento, y avanza (se mueve el punto de partida) en sentido contrario con una anchura de paso dado para buscar un mínimo local. Este proceso se repite un número de iteraciones admitido para ajustar los puntos de enfoque que el algoritmo genético dio a los helióstatos activos en primera instancia. Es importante destacar que, en este nivel, la selección de helióstatos ya no volverá a replantearse. Además, el vector gradiente se calcula de forma analítica a partir de la expresión 3 para explotar la formulación del problema y reducir el coste computacional.

La metodología aquí descrita es un resumen de la publicación disponible en [6], realizada en el seno del programa de doctorado, y donde el lector interesado puede encontrar más información.

### 2.4. Experimentación y resultados

No es posible entrar en detalles sobre la experimentación llevada a cabo y los resultados obtenidos por formar parte de artículos de revista ya publicados. No obstante, sí se resumen algunos resultados interesantes que apoyan el potencial de los métodos desarrollados.

En la figura 3 se puede ver un mapa de flujo predicho por el modelo analítico del campo (a) y el mapa equivalente al aplicar esa misma configuración en la planta real con aproximadamente 50 helióstatos (b). Nótese que se toma como realidad una simulación precisa mediante trazado de rayos. Como se puede apreciar, se trata de formas muy parecidas, especialmente teniendo en cuenta que el instante predicho no formó parte de la construcción del modelo ni tampoco la gran mayoría de los helióstatos usados. Por consiguiente, el uso del modelo del campo es fiable para predecir cómo se comporta el campo. Además, el modelo sólo tarda unas décimas de segundo en hacer la predicción mientras que el propio trazado de rayos ronda el cuarto de hora sin incluir el procesado necesario para construir la imagen.

Probada la robustez del modelo analítico para predecir el comportamiento del campo bajo demanda, se generó la base de datos paramétrica de los helióstatos de un campo para un cierto instante. Se definió entonces una forma de flujo de referencia a replicar y se lanzó el método de dos capas propuesto para ver si era capaz de configurar el campo (selección de helióstatos y puntos de enfoque) para replicarla. En la figura 4 se puede ver el resultado obtenido. Concretamente, en (a) se muestra la forma a replicar o referencia, que se formó apuntando aleatoriamente 100 helióstatos del campo. En (b) se incluye el resultado obtenido por el método de dos capas para replicarlo tal y como lo visualiza el modelo del campo. Como se puede ver, la replicación es prácticamente perfecta, por lo

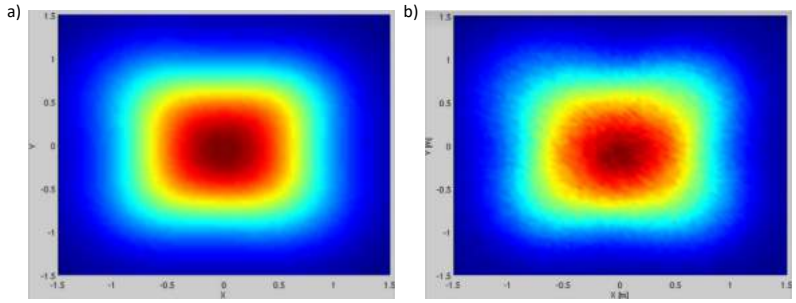



**Figura 3.** Predicción del mapa de flujo de una configuración del campo con el modelo analítico propio (a) y medición 'real' (trazado de rayos) de la misma configuración (b)

que el método fue capaz de tomar un buen conjunto de helióstatos y enfocarlos. Además, un volcado a 'realidad' (trazado de rayos preciso) como el llevado a cabo previamente confirmó la bondad de la configuración lograda por el método de dos capas para replicar la referencia (y, por extensión, reforzó la confianza en el modelo analítico).

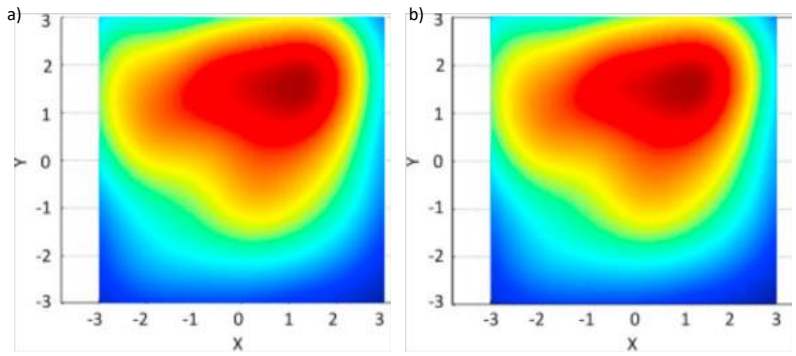



**Figura 4.** Predicción del mapa de flujo de una configuración del campo con el modelo analítico propio (a) y medición 'real' (trazado de rayos) de la misma configuración (b)

El lector interesado puede encontrar más información sobre experimentación realizada y resultados obtenidos en [7,6].

## 3. Conclusiones

En este trabajo se ha descrito una metodología transversal para replicar cualquier forma deseada sobre el receptor de una planta CSRC. Abarca desde la

selección de los helióstatos que se deben usar, hasta la asignación de los puntos de enfoque más adecuados. Su diseño se ha planteado como la resolución de un problema de optimización. Consta de dos módulos o capas: un algoritmo genético y un método de descenso por gradiente. El primero se centra en la selección de helióstatos mientras que el segundo ajusta los puntos de enfoque de todos aquellos helióstatos que se han fijado como activos. Además, se ha desarrollado una metodología para construir modelos analíticos del campo a controlar. Con ella, es posible generar una fuente de datos fiable y computacionalmente eficiente para el optimizador de dos capas. Los resultados obtenidos confirman que i) los modelos analíticos del campo son precisos, ii) pueden vincularse con el optimizador de dos capas, y iii) dicho optimizador es capaz de replicar correctamente distribuciones de flujo deseadas.

## Referencias


1. S. Alexopoulos and B. Hoffschmidt. Advances in solar tower technology. *WIREs Energy Environ.*, 6(1):1–19, 2017.
2. O. Behar, A. Khellaf, and K. Mohammedi. A review of studies on central receiver solar thermal power plants. *Renewable and sustainable energy reviews*, 23:12–39, 2013.
3. S. M. Besarati, D. Y. Goswami, and E. K. Stefanakos. Optimal heliostat aiming strategy for uniform distribution of heat flux on the receiver of a solar power tower plant. *Energy Conversion and Management*, 84:234–243, 2014.
4. M. Carasso and M. Becker. *Solar Thermal Central Receiver Systems: Performance Evaluation Standards for Solar Central Receivers*, volume 3. Springer Verlag, 1991.
5. F. J. Collado and J. Guallar. Campo: Generation of regular heliostat fields. *Renew. Energ.*, 46:49–59, 2012.
6. N. C. Cruz, J. D. Álvarez, J. L. Redondo, M. Berenguel, and P. M. Ortigosa. A two-layered solution for automatic heliostat aiming. *Engineering Applications of Artificial Intelligence*, 72:253–266, 2018.
7. N. C. Cruz, R. Ferri-García, J. D. Álvarez, J. L. Redondo, J. Fernández-Reche, M. Berenguel, R. Monterreal, and P. M. Ortigosa. On building-up a yearly characterization of a heliostat field: A new methodology and an application example. *Solar Energy*, 173:578–589, 2018.
8. A. Salomé, F. Chhel, G. Flamant, A. Ferrière, and F. Thiery. Control of the flux distribution on a solar tower receiver using an optimized aiming point strategy: Application to themis solar tower. *Solar Energy*, 94:352–366, 2013.
9. P. Schwarzbözl, R. Pitz-Paal, and M. Schmitz. Visual hflcal-a software tool for layout and optimisation of heliostat fields. In *Proceedings*, 2009.
10. W. B. Stine and M. Geyer. *Power from the Sun*. Public website: `http://www.powerfromthesun.net/book.html` (Último acceso: Mayo, 2017), 2001.
11. K. Wang and Y. L. He. Thermodynamic analysis and optimization of a molten salt solar power tower integrated with a recompression supercritical co2 brayton cycle based on integrated modeling. *Energy Conversion and Management*, 135:336–350, 2017.
12. C. J. Winter, R. L. Sizmann, and L. L. Vant-Hull. Solar power plants: Fundamentals, technology, systems. *Economics*, pages 41–53, 1991.

# Modelado y Optimización para una Gestión Eficiente de Recursos en Tecnología Termosolar

## Modelado y optimización en termosolar

Jose A. Carballo

[1] Universidad de Almería UAL
[2] CIEMAT-Plataforma Solar de Almería PSA
[3] CIESOL Centro Mixto UAL-PSA

**Resumen** El trabajo presenta la motivación de la tesis *Modelado y Optimización para una Gestión eficiente de Recursos en Tecnología Termosolar*, los objetivos principales que se plantean y el estado de desarrollo o consecución de los mismos. Además, se muestran los principales resultados y conclusiones obtenidos durante el último año gracias a los trabajos desarrollados en el marco de la tesis.

## 1. Introducción.

Uno de los principales problemas a tratar por la sociedad es la gestión racional de los recursos naturales, sobre todo fuentes de energía y agua dulce. Actualmente, numerosos organismos advierten de las ventajas ambientales, estratégicas y socioeconómicas del uso de las energías renovables frente a las energías fósiles y la gestión eficiente de los recursos.

La Unión Europea por ejemplo, considera estos hechos en gran parte de sus actividades de investigación e innovación en el Programa Marco denominado Horizonte 2020 (*H2020*). España por su parte, trata esta problemática y fija líneas estratégicas de investigación en el Plan Nacional de Investigación (Estrategia Española para la Ciencia, Tecnología e Innovación 2013-2020). El proyecto de investigación "Estrategias de control y gestión energética en entornos productivos con apoyo de energías renovables" (*ENERPRO*) se enmarca dentro de dicho plan. Este proyecto trata de analizar, diseñar y aplicar técnicas de modelado, control y optimización para conseguir una gestión eficiente de energía, agua y $CO_2$, en sistemas productivos apoyados en energías renovables y sistemas de almacenamiento.

*ENERPRO* es un proyecto formado por dos subproyectos y coordinado entre la Universidad de Almería y la Plataforma Solar de Almería a través de *CIESOL*. El subproyecto desarrollado por la Plataforma Solar de Almería (*EFFERDESAL*), en el cual se encuadra esta tesis, tiene como objetivo principal el modelado dinámico de una planta solar para desalación térmica. Esta técnica cobra especial relevancia si se tiene en cuenta que el 50 % de la población está viviendo en territorios costeros y que los avances en el campo de las renovables hacen posible la obtención de agua dulce mediante desalación con energía solar.

Por ello, en esta tesis se va a tratar la aplicación de técnicas de modelado, control y optimización para lograr una gestión eficiente de energía y agua dulce mediante el uso de energía solar. Se pretende realizar contribuciones en el campo del modelado, control y optimización de sistemas termosolares con aplicaciones industriales. De manera general se pretende estudiar, modelar, mejorar el control y optimizar, a partir de los estudios previos que muestran el gran potencial, la viabilidad técnica y económica del sistema *AQUASOL* (Fig. 1) en la configuración formada por la planta de destilación térmica acoplada a una bomba de calor de doble efecto cuya fuente de energía térmica es un campo solar.

Además, al tratarse de un proyecto coordinado, se están realizando aportaciones también en otros ámbitos de la tecnología termosolar, como es el caso de plantas solares de alta concentración, donde existen también numerosos retos a los que las técnicas de modelado y optimización pueden dar respuesta y además existen propuestas muy novedosas que pretenden aprovechar el calor residual de este tipo de instalaciones para alimentar sistemas de desalación como el tratado en esta tesis.

En resumen, con estas investigaciones se pretenden realizar aportaciones en distintos ámbitos de la tecnología termosolar mediante el uso de técnicas de modelado y optimización.

### 1.1. Instalación experimental: Sistema AQUASOL

El trabajo experimental necesario para la elaboración de esta tesis doctoral será realizado en la instalación *AQUASOL*. La primera fase de esta instalación fue construida en 1987, estaba compuesta por una desaladora multi efecto (*MED*) fabricada por *ENTROPIE*. Se pretendía probar y desarrollar el proceso de destilación acoplado con energía solar térmica. Posteriormente esta planta fue modificada dentro del marco del proyecto *AQUASOL*, que finalmente ha dado su nombre a la instalación.

El principal objetivo de la instalación es servir de banco de ensayos para probar la incorporación eficiente de la energía solar al proceso de desalación térmica (1). En esta tesis se consideraran los siguientes subsistemas de *AQUASOL*: almacenamiento térmico, una bomba de calor de absorción de doble efecto, un generador de vapor , una planta *MED* y un pequeño campo solar de canal parabólicos (campo *NEP*).

## 2. Avances y resultados.

El trabajo desarrollado en el marco de la tesis durante el año 2018, se basa en el empleo de diferentes técnicas de modelado.

En primer lugar, mediante técnicas de modelado dinámico basado en primeros principios se ha conseguido optimizar, según criterios energéticos y exergéticos, los parámetros de operación del campo de colectores solares de canal parabólico que forman parte del sistema *AQUASOL* (2). El modelo ha sido calibrado y validado con datos experimentales obtenidos en campañas de ensayos diseñadas

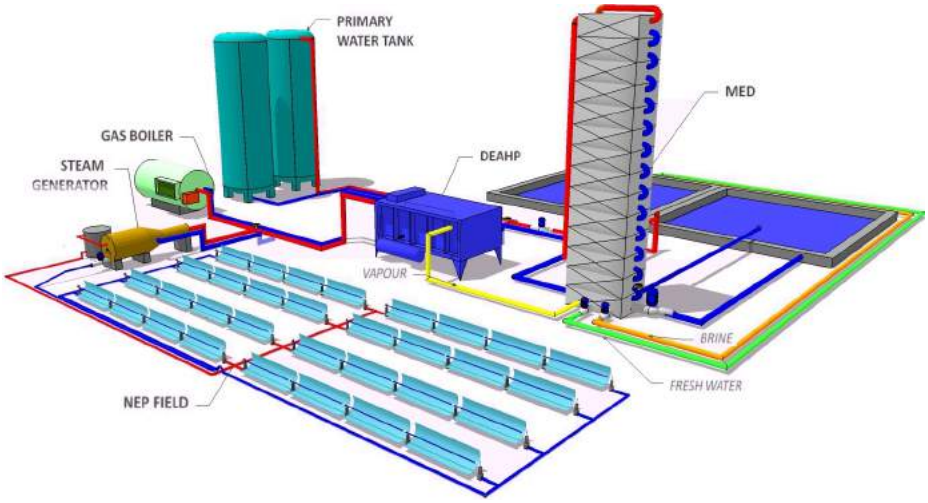



Figura 1: Sistema AQUASOL

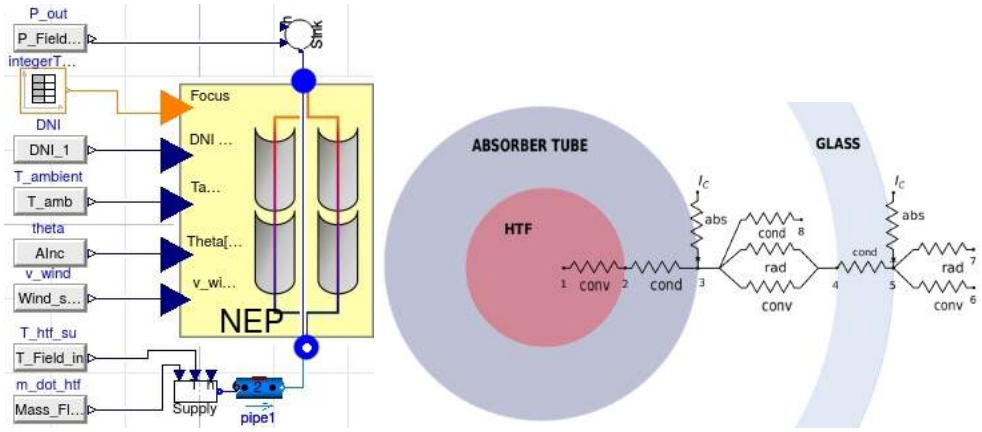
con este objetivo. También se ha trabajado en una herramienta de software multiplataforma y código abierto diseñada para estimar u optimizar los parámetros de los modelos compatibles con FMI (3).

En segundo lugar, se han empleado modelos basados en redes neuronales artificiales para desarrollar un nuevo concepto de seguidor solar. Se ha llevado a cabo una campaña experimental en campo con el objetivo de validar el concepto. Esta nueva metodología permite desarrollar nuevos esquemas de control en lazo cerrado, eliminar numerosos inconvenientes que poseen los sistemas tradicionales y aportar nueva información muy relevante para el control de los colectores y de la planta en general (4; 5).

### 2.1. Modelado dinámico basado en primeros principios y optimización.

Durante el ultimo año se ha desarrollado un modelo flexible y dinámico que permite simular el comportamiento térmico transitorio del campo *NEP* (Fig. 2a), para su uso en tareas de optimización, evaluación del rendimiento y control (2). Este modelo de la planta *NEP* se basa en los principios físicos detallados descritos en el trabajo R.Forristal (6), ver Fig. 2b, el modelo se ha adaptado, validado y mejorado con el desarrollo exergético necesario para poder realizar análisis exergéticos, como recomienda el autor. Esta nueva metodología de análisis permite definir nuevos índices de rendimiento basados en balances de exergía. Se ha desarrollado utilizando la librería *ThermoCycle* (7; 8) y se ha implementado en el lenguaje de modelado orientado a objetos basado en ecuaciones *Modelica*.

Los resultados sobre la validación del modelo muestran que este predice los valores de las variables configuradas como salidas con precisión, tanto en estado estacionario como transitorio. La propiedad de acausalidad del lenguaje de




(a) Dymola NEP model (b) Forristal Thermal model

Figura 2: NEP model

modelado hace que todas las variables se pueden configurar como entradas o salidas siempre que se establezca un sistema de ecuaciones que se pueda resolver, convirtiendo el modelo en una herramienta flexible y personalizable, que puede emplearse para la optimización, el análisis térmico o exergético y la evaluación de rendimiento en estacionario o transitorio.

Por otro lado, el análisis energético y exergético del campo de colectores de canal parabólico realizado gracias al modelo, muestra que niveles elevados de radiación solar directa (*DNI*) mejoran los índices de rendimiento térmico y exergético del campo *NEP*. Por el contrario, los cambios rápidos en los valores de *DNI* causan una reducción en ambos rendimientos. Los índices de rendimiento térmico muestran valores más altos que los índices de rendimiento exergéticos, aunque muestran un comportamiento muy similar durante la simulación. El análisis térmico revela que las principales pérdidas ocurren en la reflexión y concentración de la radiación solar, mientras que el análisis exergético revela que la destrucción de la exergía en la superficie del tubo absorbente debido a la baja temperatura de la superficie del tubo es la fuente más importante de ineficiencias exegéticas.

### 2.2. Modelado basado en redes neuronales y control.

Durante el último año también se ha trabajado en modelado basado en redes neuronales artificiales, que ha resultado en dos trabajo publicados en los que se presenta un nuevo enfoque para el control de sistemas de seguimiento solar (4; 5).

Debido a las limitaciones actuales de los sistema de seguimiento solar en cuanto a costes y problemas operacionales, se ha desarrollado un nuevo enfoque basado en visión por computador, hardware abierto de bajo coste y redes neuronales.

Los sistemas de seguimiento solar tratan de orientar el sistema de captación solar de manera óptima para maximizar el rendimiento del sistema. Para ello deben de conocer la posición relativa del Sol y el receptor. La técnica propuesta trata de detectar la posición del Sol y el área receptora mediante una cámara situada en el centro de rotación del colector solar (0) y cuyo eje óptico también es paralelo al eje óptico del colector solar ($V_A$). La cámara toma una imagen que sirve de entrada al modelo basado en redes neuronales para generar una salida con la posición en la imagen del sol y el receptor ($S'$ y $T'$). Una vez que se han identificado los elementos en la imagen, se calcula el punto medio ($A'$) entre $S'$ y $T'$ de las intersecciones del vector solar ($V_S$) y el vector receptor ($V_T$) con el plano de la cámara ($PC$). También se calcula la intersección entre $V_A$ y $PC$ ($A''$). Las diferencias entre $A'$y $A''$ se denominan error de seguimiento y se emplean como entrada para el controlador del sistema de seguimiento. Este enfoque se puede utilizar en cualquier tecnología solar, independientemente del tipo de seguidor solar, ver Fig. 3. Para probar el nuevo método se han realizado ensayos en la instalación de torre central *CESA*, puesto que el sistema de seguimiento de los colectores de estos sistemas (*Heliostatos*) son considerados los mas complejos.

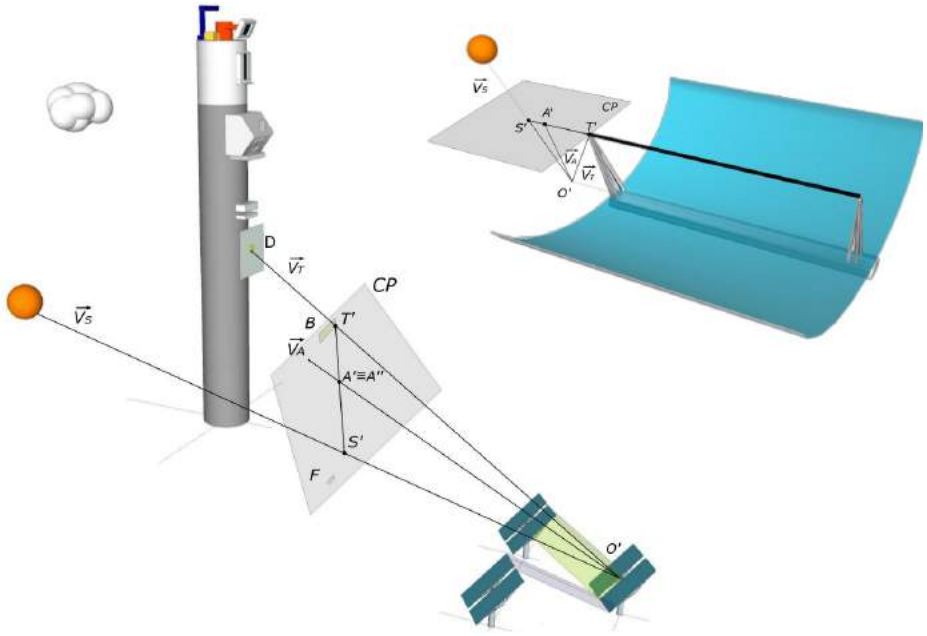



Figura 3: Nueva técnica de seguimiento solar.

La nueva metodología además de generar la señal para el control del sistema, puede proporcionar información sobre variables clave para el control avanzado del sistema de seguimiento solar, como la predicción de movimientos de nubes,

detección de bloqueos y sombras, atenuación atmosférica o medidas de radiación solar concentrada. Con todo esto se pueden mejorar las estrategias de control del sistema y por lo tanto el rendimiento del sistema.

También se ha trabajado en la implementación de este nuevo enfoque en código abierto. Concretamente se ha empleado Tensorflow (9). Esta librería hace que la implementación sea más flexible y aumenta las capacidades de desarrollo (10).


Figura 4: Captura de pantalla del sistema de control con Tensorflow sobre el campo CESA

Las pruebas preliminares llevadas a cabo con éxito en la Plataforma solar de Almería (*PSA*), en campo real (Fig. 4) como en laboratorio (Fig. 5), revelan el gran potencial y muestran el nuevo enfoque como una buena alternativa a los sistemas tradicionales.

También se ha participado en diferentes actividades divulgativas como *"la Noche Europea de los Investigadores 2018"* (Fig. 6) o la *Semana de la Ciencia*.

## 3. Conclusiones y trabajos futuros.

Se ha desarrollado y validado un modelo que se ha empleado posteriormente para realizar un estudio de caracterización del comportamiento térmico y exergético del campo *NEP* tanto en estado estacionario como transitorio. El estudio ha sido publicado en una revista del primer cuartil (2). El trabajo desarrollado en la herramienta de simulación de modelos *FMI* fue presentado en

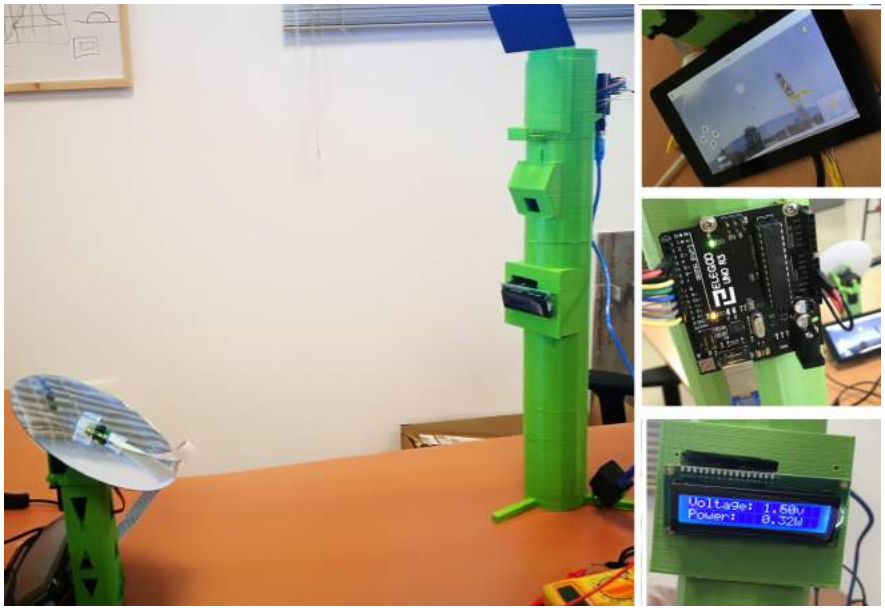



Figura 5: Prototipo funcional construido mediante impresión 3D.


Figura 6: Actividad divulgativa *Noche Europea de los investigadores*.

congreso (3). Las conclusiones principales del trabajo son que el modelo se ajusta a los requerimientos de diseño y que el análisis térmico y exergético de la instalación apunta a causas diferentes de ineficiencias. Este modelo resulta ser flexible, preciso y altamente configurable, lo que permite adaptarlo fácilmente y le otorga un amplio campo de aplicación.

Actualmente se está trabajando en el desarrollo del modelo del generador de vapor para integrarlo con el modelo del campo *NEP* y otro modelo simplificado de la *DEHP*. Finalmente se integrarán todos los sistemas y se procederá a la validación y la optimización del sistema completo (sistema $AQUASOL$).

En cuanto al modelado basado en redes neuronales, el desarrollo del nuevo enfoque ha dado como resultado dos publicados en revistas del primer cuartil (4; 5) y una contribución a congreso (10). Además, la técnica descrita se encuentra bajo proceso de patente. De acuerdo con los resultados obtenidos en los ensayos preliminares, se puede concluir que el nuevo enfoque propuesto para los sistemas de seguimiento solar es válido, completamente funcional y muestra un amplio margen de mejora. El nuevo enfoque es independiente de la tecnología solar, el tamaño del sistema, la ubicación y el tiempo. No se ve afectado por errores de apunte como la inclinación del pedestal, las cargas de viento o la posición aparente del Sol. Además, el enfoque propuesto proporciona otras ventajas como la capacidad de detección de nubes, bloques y sombras, atenuación atmosférica o medición de radiación solar concentrada, que puede mejorar las estrategias de control del sistema y, por lo tanto, el rendimiento del sistema. El nuevo enfoque combinado con las técnicas de control tradicionales hace posible los esquemas de control de lazo cerrado.

Los trabajos futuros incluyen probar estos métodos y algoritmos, realizar una campaña de ensayos para realizar un entrenamiento mas profundo del modelo basado en redes neuronales con el objetivo de mejorar aún más los resultados obtenidos y reducir el costo computacional. Otra tarea importante es el control automático de los colectores solares mediante el nuevo método de seguimiento presentado.

## 4. Acrónimos

| | |
|---|---|
| A' | Punto de apunte |
| A" | Punto ideal de apunte |
| CIESOL | Centro mixto UAL-PSA |
| DEAHP | Double effect absorption heat pump |
| DNI | direct normal irradiance |
| EFEDESAL | Efficient energy control and management of solar thermal desalination systems |
| ENERPRO | Estrategias de control y gestión energética |
| FMI | functional mock-up |
| MED | multi effect desalination |
| PC | Plano de la cámara |
| PSA | Plataforma soalr de Almería |
| S' | Posición del sol |
| T' | Posición de la diana |
| UAL | Universidad de Almería |
| Va | Vector de apunte |
| Vs | Vector solar |
| Vt | Vector de diana |

# Bibliografía


[1] Alarcon-Padilla, D.C., Garcia-Rodriguez, L., Blanco-Galvez, J.: Assessment of an absorption heat pump coupled to a multi-effect distillation unit within AQUASOL project. Desalination **212**(1-3) (2007) 303–310

[2] Carballo, J.A., Bonilla, J., Berenguel, M., Palenzuela, P.: Parabolic trough collector field dynamic model: Validation, energetic and exergetic analyses. Applied Thermal Engineering **148** (2019) 777–786

[3] Bonilla, J., Carballo, J.A., Roca, L., Berenguel, M.: Development of an open source multi-platform software tool for parameter estimation studies in fmi models. In: Proceedings of the 12th International Modelica Conference, Prague, Czech Republic, May 15-17, 2017. Number 132, Linköping University Electronic Press (2017) 683–692

[4] Carballo, J.A., Bonilla, J., Roca, L., Berenguel, M.: New low-cost solar tracking system based on open source hardware for educational purposes. Solar Energy **174** (2018) 826–836

[5] Carballo, J.A., Bonilla, J., Berenguel, M., Fernández-Reche, J., García, G.: New approach for solar tracking systems based on computer vision, low cost hardware and deep learning. Renewable energy (2018)

[6] Russell, F.: Heat Transfer Analysis and Modeling of a Parabolic Trough Solar Receiver Implemented in Engineering Equation Solver. National Renewable Energy Laboratory (October) (2003) 164

[7] Quoilin, S., Desideri, A., Wronski, J., Bell, I., Lemort, V.: ThermoCycle: A Modelica library for the simulation of thermodynamic systems. (2014) 683–692

[8] Desideri, A., Hernandez, A., Gusev, S., van den Broek, M., Lemort, V., Quoilin, S.: Steady-state and dynamic validation of a small-scale waste heat recovery system using the thermocycle modelica library. Energy **115** (2016) 684–696

[9] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI. Volume 16. (2016) 265–283

[10] Carballo, J.A., Bonilla, J., Berenguel, M., Fernandez-Reche, J., Garcia, G.: Machine learning for solar trackers. In: Proceedings of the International SolarPaces, Casablanca, Morocco, October 2-5, 2018. (2018)

# A Recommender System for Smart User Interfaces using Machine Learning and Microservices

Antonio Jesus Fernández-García

Applied Computing Group. University of Almeria, Spain.
ajfernandez@ual.es

**Abstract.** The burst of artificial intelligence and, more specifically, machine learning, is contributing to change the way that humans manage or communicate with information systems and with their environment. The use of machine learning has already been widely applied by large companies and it is expanding to the general population, spreading to each of the parts of our lives, creating personalized recommendations for us that even we ourselves do not know. In this thesis, we establish a methodology that makes use of machine learning technologies to create recommendation models. To illustrate all the processes and to describe how the proposed methodology can be applied a a real scenario, we present as a case study a recommendation system that allows the creation of smart user interfaces. We make use of the users' behavior to make user interfaces that evolve over time according to the users' needs.

## 1 Introduction

The progress of the computers processing capacity and the cloud computing payment models that make computing affordable to (almost) everyone have made it possible to manage a huge amount of data that was unimaginable until only a few years ago. This computational progress has enabled us to run artificial intelligence algorithms that, although they were already developed mathematically years ago, have now been improved and democratized and therefore, they are able to be executed on a large scale. It affects to a large variety of fields. Accurately forecasting, precisely identifying trends and the discovery of behavior patterns clearly optimize resource usage or consumption as well as generate new knowledge in science and research facilities; enabling faster and better decisions in politics, retail, weather, sport, science, research, real estate, sports or healthcare among many others fields.

## 2 Objectives and Motivation

The main objective of this research work is to present a strategy to develop smart component-based interfaces by creating recommender systems using data

analysis techniques. To illustrate the whole process we will focus on the creation of a recommender system that assists users' in component-based applications (mashups) at finding the most suitable component for them at any time in any situation as a first step to create smart component-based user interfaces.

We pretend to establish a methodology that makes use of machine learning technologies to create recommendation models. In order to do that, there are certain aspects that need to be considered: 1) It is necessary to have a good insight of the context where the recommendation system will be deployed; 2) It is necessary to define an effective and efficient strategy to acquire the necessary data by creating data acquisition systems; and 3) It is necessary to have a thorough knowledge of machine learning algorithms, feature selection methods [5] and feature engineering techniques [12] to really get some insight from data. Although these processes require a high degree of dependence from the context where they will be applied, we have structured and documented the methodology in such a way that it is extensible to be applied in every situation where a recommendation system needs to be deployed.

To illustrate all the processes and to describe how the proposed methodology can be applied in a real scenario, we present as a case study a recommendation system that allows the creation of smart user interfaces. Recently, a large number of fields make use of computational intelligence methods to make predictions based on the users' behavior and, in the case study, we will make use of the users' behavior to make user interfaces that evolve over time according to the users' needs based on their previous behavior. The real scenario where we apply the methodology is called ENIA[1] [6]

The user interface of the case study is a component-based (mashup) graphical user interface. This is a popular kind of interface where users customize their own interface from repositories that contain a large number of components at their disposal. To work in that area, it is imperative to understand the morphology of these kinds of interfaces in the first place. For that reason, we formally describe their structure as well as understand the main operations that can be carried out in them and how users perform such operations, including the implications that they may have in the information system behavior.

After that, we present a flexible data acquisition system capable of capturing the human-computer interactions performed by users over mashup interfaces with the aim of storing them. To achieve that purpose, an architecture of microservices has been designed in the cloud to detect, acquire, and collect the interactions performed over this kind of interfaces. The whole process is ready for acquiring internal data of the information system as well as context information and location awareness. To validate the data acquisition system, some tests on empirical case studies have been developed. Efficiency and effectiveness have also been determined by evaluating the performance of the acquisition system during different load tests. In addition, in order to ensure the software quality, a continuous integration strategy for software development and an easy manage-

[1] ENIA (ENvironmental Information Agent): http://acg.ual.es/projects/enia/ui/

ment of the code have been used, facilitating the software maintenance alongside the microservice architecture, where functionalities are well encapsulated [3].

Finally, we make predictions and recommendations to end users by creating a recommender system using intelligent data analysis methods. Once the interaction data required is gathered and a dataset is built, we address the problem of transforming the original dataset to an optimized dataset ready to be used in machine learning algorithms. The transformation is made through the application of feature engineering techniques and feature selection methods. Moreover, many aspects, such as contextual information, the use of the application across several devices with many forms of interaction, or the passage of time (components are added or removed over time), are taken into consideration. Once the dataset is optimized, a series of experiments are conducted applying several machine learning algorithms to the optimized dataset (before and after applying feature selection methods) to create recommendation models. Lastly, to determine which recommender model offers a better performance, several metrics are used to evaluate them [4]. All the experiments have been carried out using Microsoft Azure Machine Learning Studio [8].

Once we have created the best possible recommendation model, we pay attention to its deployment on real mashup applications. This process is not always easy and many issues may arise such as conflicts between the components instantiated. Our methodology creates a web service that encapsulates the recommendation model and produces personalized recommendations in real-time based on the users' behavior. It incorporates a conflicts management module that determines the suitability of the recommendations before being offered to end users.

Thus, through the deployment of the recommendation systems that have better results and led to no conflicts or breach any constraint, it is possible to offer customized suggestions created exclusively to users in component-based applications, enhancing their user experience and the application engagement.

Figure 1 graphically shows the big picture that it aims to be achieved in this doctoral thesis which integrates the knowledge areas referred above; a framework scenario which, through the user interaction with the interface and the context information, allows the creation of intelligent, dynamic and evolving user interfaces providing recommendations to users' or modifying the user interface structure at real time. This is an important extension of the work developed by Criado et al [1, 2].

## 3 Results and Conclusions

In this dissertation, we address the problem of creating recommender systems that are able to improve the users experience in mashup component-based applications by forecasting their needs in concrete situations. By doing that, we pretend to optimize the possibilities of successfully achieving a good position of software applications in the increasingly competitive software market.

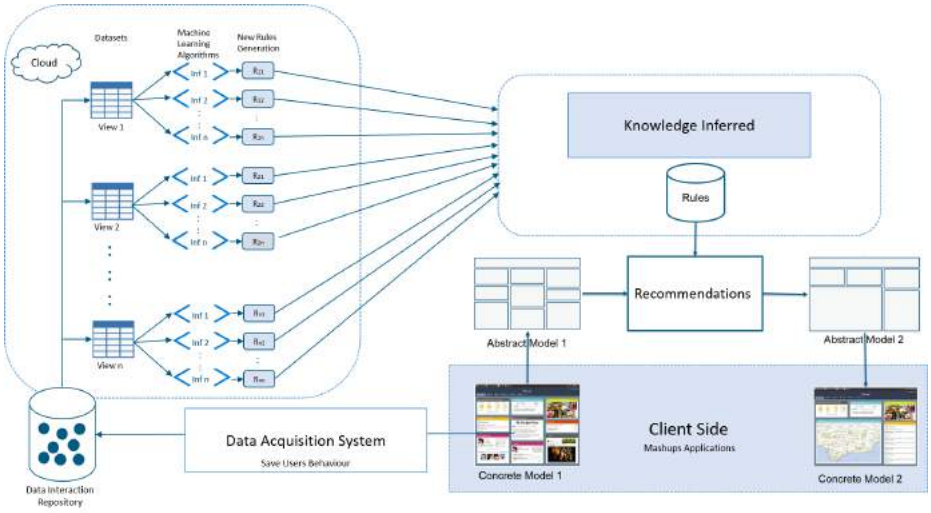



**Fig. 1.** Entire Perspective

Previously, we graphically illustrated in Figure 1 the entire perspective of the work we pretended to develop in this dissertation. Now, in Figure 2 we can summarize our contribution with a picture that reflects the main work developed. As this figure shows, we can divide the whole work into three main parts: the data acquisition system, the microservices architecture for creating datasets and the creation and deployment of the recommendation models created. We have structured the results and conclusions according to these three main parts to easily visualize the contribution of this research work.

We have presented a data acquisition strategy that allows mashup user interfaces to store the interaction that users perform over them into a relational database. Tthe work presented is the first in the *Mashup User Interfaces* literature that addresses the problem of capturing the human-computer interactions performed by users over mashup user interfaces through a flexible data acquisition system with the aim of storing such interactions in a relational database. Our work reports the following R&D activities conducted in this regard:

— We have deeply investigated the mashup UI morphology and we have proposed a database schema to store the interaction based, not only on the mashup morphology, but also on the information managed by these information systems, according to their users and services.
— A relational database for storing the interaction has been design under the consideration that none a nuance of the data that is involved in an interaction is missed and it can be easily accessible for further purposes.
— A microservice architecture has been designed and a data acquisition system has been deployed in order to acquire data from mashup applications whose user interfaces run distributed across multiple devices.

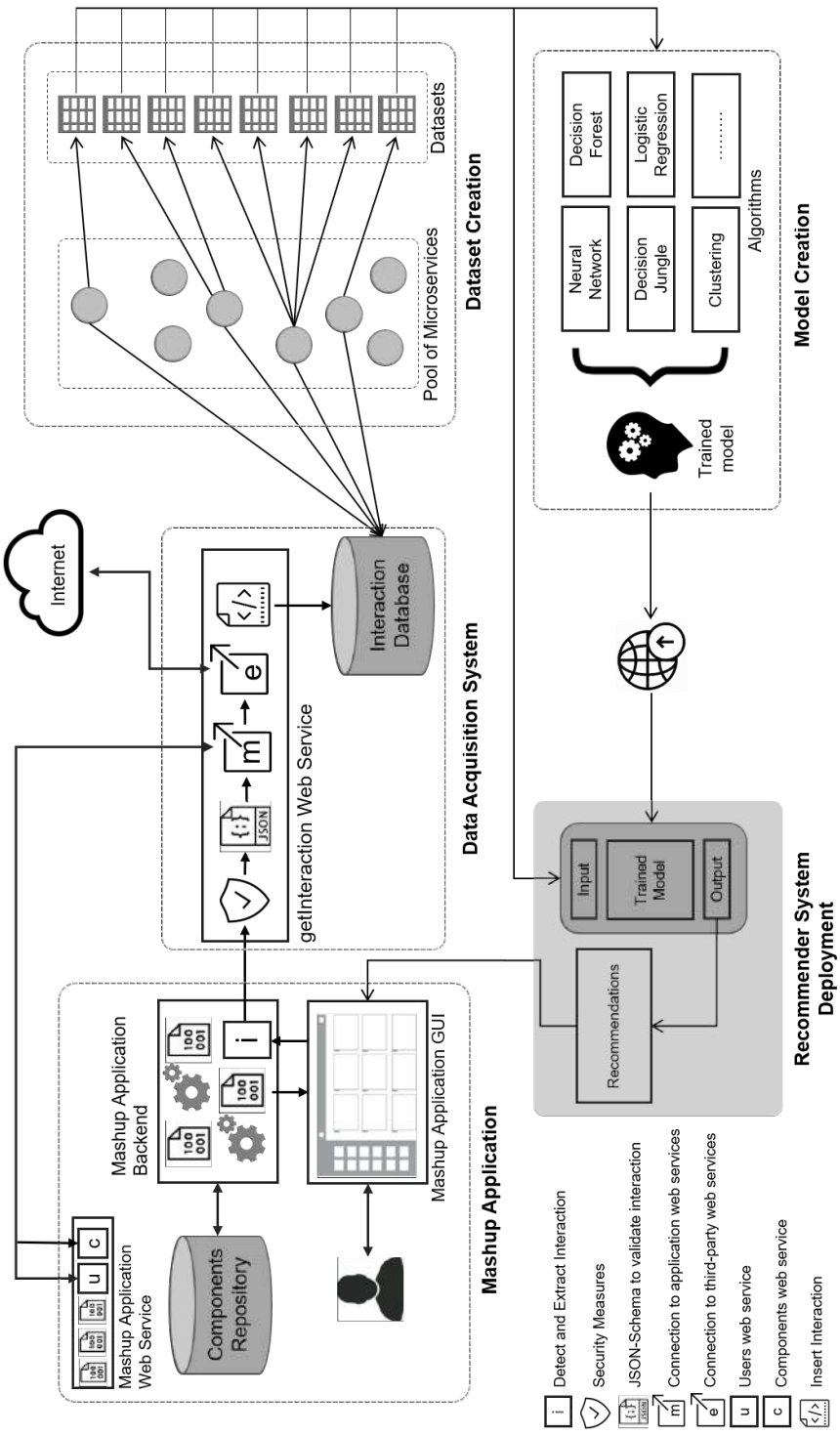



**Fig. 2.** Figure that summarizes the dissertation contribution

— Different user profiles and categories can be handled as well as there can be guest users. This classification significantly influences on predictive models.
— It can handle the context awareness capability of some devices in ubiquitous computing system which it is interesting to enhance the user experience.
— Users accessing simultaneously to the component-based application through multiple devices using concurrent sessions is supported.
— Multiples forms of interaction are also supported. This is important due to the fact that each device can present a different form of interaction and even some devices can have several of them.
— It is ready to work with new devices that incorporate Natural User Interfaces (NUIs) since the device and form of interaction are taken into consideration.

Once the data acquisition system has done its work, the interaction data has to be obtained in form of dataset to apply data analysis techniques from the relational database. Usually, the interaction data fetched is too raw and, in order to create optimal datasets for further analysis with machine learning algorithms, a deep feature engineering work is performed.

The new optimized datasets increase the possibilities of creating accurate prediction models that might help to improve the user experience when using the mashup application through heterogeneous devices. This is done by stepping to the users' needs and providing them with a customized user experience. For this customization, in addition to the user behavior, it is considered both, the user device and the form of interaction of each interaction in a concrete moment along with information from the location and the environment. Thus, the user interface customization is carried out according to the study of the users' behavior, their devices, and the context. Our work reports the following R&D activities conducted in this regard:

— Each microservice is service-oriented to facilitate an agile development as well as an easy software test, inspection and maintenance.
— The microservice granularity allows us to easily design, implement and deploy microservices that create datasets with different purposes that easily escalate according to their needs.
— The microservice-based architecture allows developer teams to simultaneously implement new functionalities independently. Also, each team can use the best possible technologies in each case to solve the problem and they should easily be integrated into the system.
— Microservices are equipped with REST API web services for exposing their functionalities to other microservices, receiving feedback from algorithms or other sources, and communicate with third-party clients.
— The process of creating datasets by the proposed architecture is autonomous and continuous, hence, datasets are always updated with the latest interaction data.
— Datasets are always updated thanks to an automated system that periodically generates new datasets. Also, datasets can be generated on demand by calling a web service in each microservice.

Finally, we address the problem of creating a recommender system that is able to suggest to users of component-based interfaces which are the most suitable components for them to use at a specific time. By forecasting the component most closely aligned to each situation, we aim to improve the user experience in the software application and thus, optimize the possibilities of successfully achieving a good position in the increasingly competitive software development market. Our work reports the following R&D activities conducted in this regard:

— We have applied feature selection methods that support all data type features such as the *Mutual Information Score* and *Chi-Squared Statistic* to determine the relevance of each feature of the dataset and to create subsets that contain the most significant features.
— We have created several models using a dataset that contains the interactions performed by users in component based applications after applying feature engineering techniques and feature selection methods.
— We have created the recommendation models parameterizing some well known classification algorithms such as *Multiclass Decision Forest* [9], *Multiclass Decision Jungle* [11], *Multiclass Neural Networks* [10] and *Multiclass Logistic Regression* [7].
— We have evaluated and compared the models created in terms of overall accuracy and average accuracy.
— We have built and analyzed a *Confusion Matrix* that offers the specifying accuracy of each model created, including the deviation that the model may have predicting a class compared with the labeled class.
— We have created a web service that encapsulates the trained recommender model, accepts as an input real-time interaction data and produces recommendations.
— We have incorporated a conflicts management module that manages potential conflicts that may be raised by checking if the recommendation does not contradict any other recommendation or violates any user interface constraint.
— We have created a decision system to decide whether a recommendation should be finally suggested to end users or not by analyzing the suitability of the recommendation and the output of the conflicts management module.

## References

1. Javier Criado, Diego Rodríguez-Gracia, Luis Iribarne, and Nicolás Padilla. Toward the adaptation of component-based architectures by model transformation: behind smart user interfaces. *Software: Practice and Experience*, 45(12):1677–1718, 2015.
2. Javier Criado, Cristina Vicente-Chicote, Nicolás Padilla, and Luis Iribarne. A model-driven approach to graphical user interface runtime adaptation. In *5th International Workshop on Models*, pages 49–59, 2010.
3. Antonio Jesús Fernández-García, Luis Iribarne, Antonio Corral, Javier Criado, and James Z. Wang. A flexible data acquisition system for storing the interactions on mashup user interfaces. *Computer Standards & Interfaces*, 59:10 – 34, 2018. DOI: 10.1016/j.csi.2018.02.002.

4. Antonio Jesús Fernández-García, Luis Iribarne, Antonio Corral, Javier Criado, and James Z. Wang. A recommender system for component-based applications using machine learning techniques. *Knowledge-Based Systems*, 164:68–84, 2019. DOI: 10.1016/j.knosys.2018.10.019.
5. Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16 – 28, 2014.
6. L. Iribarne. ENIA Project. Environmental Information Agent. `http://acg.ual.es/projects/enia/`, 2016. Online; last accessed January 2019.
7. P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.
8. Microsoft Corporation. Microsoft Azure Machine Learning Studios. https://studio.azureml.net. Online; last accessed 18 December 2017.
9. John Ross Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986.
10. Raúl Rojas. *Neural Networks: A Systematic Introduction*. Springer-Verlag New York, Inc., New York, NY, USA, 1996.
11. Jamie Shotton, Toby Sharp, Pushmeet Kohli, Sebastian Nowozin, John Winn, and Antonio Criminisi. Decision jungles: Compact and rich models for classification. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 234–242. Curran Associates, Inc., 2013.
12. A. Zheng and A. Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Incorporated, 2018.

# Tratamiento de Spatial Big Data: Particionado, Métricas y Consultas

Francisco José García García
E-mail: paco.garcia@ual.es

Universidad de Almería

**Abstract.** El procesamiento de consultas sobre datos espaciales ha atraído la atención de los investigadores durante décadas, debido al amplio rango de aplicaciones que utilizan información espacial que han ido apareciendo y al gran volumen de datos que muchas de éstas manejan. En base a ello, en esta tesis se han analizado e implementado algoritmos paralelos y distribuidos para dos de las consultas de join basadas en distancias más estudiadas, K Closest Pair Query ($k$CPQ) y K Nearest Neighbours Join Query ($k$NNJQ), sobre dos de los sistemas de gestión de datos espaciales distribuidos más estudiados en la actualidad (SpatialHadoop y LocationSpark). También se han presentado nuevas técnicas de particionado basadas en diagramas de Voronoi) que junto con nuevas métricas y propiedades mejoran el rendimiento de dichas consultas. Además, se ha presentado el trabajo realizado sobre una nueva consulta denominada Group K-Nearest Neighbor (GNN) en un entorno distribuido como es SpatialHadoop. Por último, se presenta un trabajo en curso que busca mejorar nuestra consulta Reverse K-Nearest Neighbor (R$k$NN) en MapReduce mediante el uso de técnicas de poda de particiones (SLICE). Todos los algoritmos y técnicas han sido evaluados mediante experimentos con grandes conjuntos de datos reales en diferentes escenarios.

**Keywords:** Big Data · MapReduce · Spark · Spatial Data Processing · Spatial Query Evaluation · SpatialHadoop · LocationSpark · Distance Joins · Partitioning · Voronoi-Diagram · GNNQ · RNNQ.

## 1 Introducción

Estamos viviendo en la era del Big Data, y los datos espaciales no son una excepción. Aplicaciones móviles, automóviles, dispositivos GPS, barcos, aviones, dispositivos IoT, etc. están generando cantidades ingentes de datos con características espaciales. Esto ha motivado la necesidad de desarrollar sistemas para Big Data espacial que utilicen tecnologías novedosas para el procesamiento de datos espaciales a gran escala en clústeres de ordenadores de un entorno distribuido. Estos Sistemas de Gestión de Datos Distribuidos (Distributed Data Management Systems or DDMSs) se clasifican en basados en disco [11] o basados en memoria principal [22].

Los Sistemas de Gestión de Datos Espaciales Distribuidos (DSDMSs) basados en disco se caracterizan por estar basados en Hadoop y los más representativos

son SpatialHadoop [4] y Hadoop-GIS [2]. Por otro lado, los DSDMSs basados en memoria se caracterizan por estar basados en Spark y los más importantes son GeoSpark [21] y LocationSpark [17]. Estos sistemas permiten a los usuarios trabajar con datos espaciales distribuidos sin preocuparse sobre la computación distribuida y la tolerancia a fallos.

El procesamiento eficiente de consultas espaciales para grandes volúmenes de datos espaciales es un desafío y como se muestra en [4], se están proponiendo distintos algoritmos que intentan dar respuesta a diferentes consultas, como rango, $k$NN, spatial joins y skyline queries, en este tipo de entornos distribuidos. Por lo tanto, el principal objetivo de esta tesis doctoral es, a partir del estudio y uso de diferentes entornos paralelos y distribuidos, desarrollar nuevos algoritmos eficientes que aprovechen las características que proporcionan dichos entornos. La eficiencia y escalabilidad de nuestras propuestas se demostrarán con los resultados de la ejecución de un extenso conjunto de experimentos sobre conjuntos de datos, tanto sintéticos como reales.

La tesis se enmarca dentro del proyecto TIN2017-83964-R del Ministerio de Economía y Competitividad (MINECO) del Gobierno de España.

El resto de este artículo continua con la presentación de los avances realizados durante el último año en el desarrollo de la tesis. Finalmente, se muestran las conclusiones obtenidas hasta el momento y se muestra una descripción de posibles trabajos futuros a realizar.

## 2 Avances

### 2.1 Efficient Distance Join Query Processing in Distributed Spatial DataManagement Systems

En este artículo, que actualmente se encuentra en revisión, extendemos nuestros anteriores trabajos [9] y [6] mediante la implementación de nuevos algoritmos ($k$NNJQ y $\varepsilon$DRJQ), su mejora con el uso de técnicas de reparticionado de las zonas espaciales más densas, la extensión al uso de objetos espaciales complejos (polígonos) y un estudio experimental de los diferentes algoritmos con conjuntos de datos reales en SpatialHadoop y LocationSpark.

Primero presentamos nuevos algoritmos en MapReduce para $k$NNJQ y $\varepsilon$DRJQ. La consulta $k$NNJQ, dados dos conjuntos de puntos ($\mathbb{P}$ y $\mathbb{Q}$) y un número positivo $k$, encuentra para cada punto de $\mathbb{P}$, sus $k$ vecinos mas cercanos en $\mathbb{Q}$. Un ejemplo de esta consulta consistiría en encontrar los 5 ($k$) parques más cercanos para cada lago. Otro ejemplo de la misma podría ser, para cada casa (punto), encontrar las 3 ($k$) áreas (polígono) con alta criminalidad más cercanas. Nuestra propuesta de algoritmo para entornos distribuidos, basada en [13], consta de las siguientes fases:

- **Bin $k$NNJ**, que consiste en un Bin-Spatial Join [23] donde el operando de join es el $k$NN.
- **$k$NNJ en las celdas solapadas**, que aplica $k$NN en aquellas celdas que solapan con las listas $k$NN parciales de la fase anterior.

– **Combinación de resultados**, de las listas $k$NN de las dos fases anteriores.

La consulta $\varepsilon$DRJ, dados dos conjuntos de puntos ($\mathbb{P}$ y $\mathbb{Q}$) y un límite de distancia $\varepsilon$ encuentra para cada punto $p_i \in \mathbb{P}$, todos los puntos en $\mathbb{Q}$ que se encuentra en la forma circular, centrada en $p_i$ con radio $\varepsilon$. Un ejemplo de $\varepsilon$DRJ puede ser encontrar todos los parques en 5 km ($\varepsilon$) para cada lago. La implementación propuesta se basa en la presentada para el algoritmo $k$NNJQ y debido a que la distancia $\varepsilon$ es conocida, esta consiste en la combinación de las fases *Bin kNNJ* y *kNNJ en las celdas solapadas* en una sola fase. Por lo tanto, estamos hablando de un Reduce-based Join ya que la última fase de *Combinación de resultados* no es necesaria.

Otro aporte de este trabajo, es la extensión de los algoritmos DJQ en MapReduce para el procesado de objetos espaciales complejos. Normalmente los conjuntos de datos reales no se limitan solo a puntos, sino que incluyen otros objetos geométricos, como segmentos de línea, polígonos, regiones, etc. De hecho, un conjunto de datos que contenga los edificios de una ciudad puede usar polígonos, mientras que las carreteras pueden ser representadas con segmentos de línea. Para extender nuestros algoritmos, hay que modificar cada una de las fases que los constituyen. Inicialmente, tenemos que tener en cuenta que el particionado en SpatialHadoop replica cada geometría en todos las particiones con las que interesecciona. Por ello, para eliminar posibles resultados duplicados, se ha utilizado la técnica *reference-point duplicate avoidance technique* [4], la cual consiste en seleccionar un único punto de la geometría y descartar sus apariciones en las particiones en las que dicho punto no esta presente. Además, para simplificar las operaciones y cálculos de distancias en los algoritmos, estos se realizaran sobre el MBR (Minimum Bounding Rectangle) que cubre los diferentes objetos espaciales. De esta manera, los algoritmos basados en barrido del plano solo tienen que calcular la mínima distancia entre MBRs sin realizar cálculos complejos basados en sus formas, los cuales solo se realizarán en la fase final de las consultas.

Cuando se realizan tareas MapReduce, un problema que normalmente aparece es el de datos sesgados (*skewed data*). De manera general, este problema consiste en que algunas particiones contienen un número mayor de elementos que el resto y por lo tanto provocan que algunas tareas tarden mucho más que el resto y se produzca un retraso en la obtención del resultado final. Para solucionar el anterior problema se han propuesto una serie de mejoras, que utilizando datos ya particionados por SpatialHadoop (ej., Grid, Quadtree), reparticionan aquellas particiones que exceden un cierto número de elementos. Esta *técnica de reparticionado* se usa principalmente para $k$NNJQ y $\varepsilon$DRJQ en SpatialHadoop, aunque puede ser utilizadas con otros algoritmos DJQ y DSDMSs. En nuestra propuesta se han implementado dos tipos de técnicas de reparticionado. Por un lado, *Reparticionado basado en Grid*, que dado un número máximo de elementos $L$, divide la partición original en tantas *filas* y *columnas* como sea necesario para que cada celda tenga como mucho $L$ elementos. La principal ventaja es ésta técnica de reparticionado es que no se necesita ningún procesamiento previo para dividir la partición en sub-particiones. Por otro lado, una técnica de

*reparticionado basado en Quadtree* ha sido diseñada e implementada. Debido a que este método de reparticionado se basa en como los datos están distribuidos, es necesario realizar una tarea previa a los algoritmos del DJQ.

Por último, se ha realizado un estudio de rendimiento detallado que demuestra que los algoritmos DJQ propuestos son eficientes, robustos y escalables respecto a los diferentes parámetros utilizados (tamaño de datos, $k$, $\varepsilon$, número de nodos de computación ($\eta$), etc.). Además se ha demostrado que LocationSpark es el ganador en tiempo de ejecución cuando se combinan conjuntos de tamaño mediano, debido a la eficiencia de su procesado en memoria que proporciona Spark. Sin embargo, SpatialHadoop es más rápido con los conjuntos de datos más grandes, debido a su madurez y robustez, debido al mayor tiempo dedicado para su investigación y desarrollo.

### 2.2 Voronoi-Diagram Based Partitioning for Distance Join Query Processing in SpatialHadoop

Las principales contribuciones de este artículo [8] son el diseño e implementación de una técnica de particionado eficiente basada en Diagramas de Voronoi en SpatialHadoop, su aplicación para la mejora de nuestros algoritmos $k$CPQ [9] y $k$NNJQ (sección anterior) en MapReduce y la ejecución de un extenso conjunto de experimentos que estudian su eficiencia y escalabilidad comparada con otros algoritmos existentes en SpatialHadoop.

Para su definición, sea $\mathcal{R} = \{r_0, r_1, \cdots, r_{r-1}\}$ un conjunto de $r$ puntos distintos en el plano; llamados generadores o *pivotes*. Podemos definir el *Diagrama de Voronoi* de $\mathcal{R}$ como la subdivisión del plano en $r$ celdas, una para cada pivote en $\mathcal{R}$, con la propiedad de que cada punto $p$ esta en la celda correspondiente a un pivote $r_i$ si y solo si $dist(p, r_i) < dist(p, r_j)$ para cada $r_j \in \mathcal{R}$ con $j \neq i$. Podemos denotar el Diagrama de Voronoi generado por $\mathcal{R}$ como $VD(\mathcal{R})$. La celda de $VD(\mathcal{R})$ que corresponde al pivote $r_i$ se llama Celda Voronoi de $r_i$ y se denota como $VC(r_i)$. El Diagrama de Voronoi tiene también la siguiente propiedad: $VD(\mathcal{R}) = \bigcup_{i=0}^{r-1} VC(r_i)$ y $\bigcap_{i=0}^{r-1} VC(r_i) = \emptyset$.

Según [12], dado un conjunto de datos $\mathbb{P}$, la técnica de *particionado basado en Diagramas de Voronoi* consiste en seleccionar un conjunto $\mathcal{R}$ de pivotes (que no tiene que pertenecer necesariamente a $\mathbb{P}$) como *pivotes*, y entonces dividir los puntos de $\mathbb{P}$ en $|\mathcal{R}|$ particiones disjuntas, donde cada punto se asigna a la partición del pivote más cercano $r_i$. En el caso de existir múltiples pivotes asignados a un punto particular, entonces este punto se asigna a la partición con el menor número de puntos. De esta forma, todo el espacio de datos se divide en $|\mathcal{R}|$ Celdas de Voronoi disjuntas. Sea $\mathcal{R}$ el conjunto de pivotes seleccionado, $\forall r_i \in \mathcal{R}$, $\mathcal{P}_i^{\mathbb{P}}$ denota el conjunto de puntos de $\mathbb{P}$ que tiene a $r_i$ como su pivote más cercano. Además, denotamos $U(\mathcal{P}_i^{\mathbb{P}})$ y $L(\mathcal{P}_i^{\mathbb{P}})$ como la máxima y mínima distancia desde el pivote $r_i$ a los puntos de $\mathcal{P}_i^{\mathbb{P}}$, respectivamente. Esto es, $U(\mathcal{P}_i^{\mathbb{P}}) = \max\{dist(p, r_i) : \forall p \in \mathcal{P}_i^{\mathbb{P}}\}$ y $L(\mathcal{P}_i^{\mathbb{P}}) = \min\{dist(p, r_i) : \forall p \in \mathcal{P}_i^{\mathbb{P}}\}$.

Para incorporar esta técnica en SpatialHadoop se han implementado los siguientes pasos: (1) Inicialmente, SpatialHadoop proporciona una muestra aleatoria,

$\mathcal{S}$, de un conjunto de datos $\mathbb{P}$ y los valores de los parámetros $x$ (número de particiones basado en el tamaño de fichero y la capacidad de bloque HDFS) y $s = |\mathcal{S}|$ (*Sampleado*). (2) Un conjunto $\mathcal{R}$ de *pivotes* se obtiene de la muestra aleatoria $\mathcal{S}$ (*División del espacio*), usando alguna técnica de selección de pivotes como *selección aleatoria* ($Voronoi_R$) o un algoritmo *k-means* ($Voronoi_k$) tal y como se describe en [12]. Para el primero, se generan $\lfloor s/x \rfloor$ conjuntos aleatorios de pivotes y se elige el que muestra la mayor suma de distancias para cada par de pivotes. Para el segundo, un algoritmo $k$-means estándar se inicializa usando un conjunto aleatorio de $x$ pivotes de la muestra aleatoria $\mathcal{S}$ y para reducir el tiempo, cuando se particiona un elevado numero de elementos, se utiliza una distancia límite como criterio de convergencia. (3) Finalmente, los puntos son asignados a su pivote más cercano $r_i \in \mathcal{R}$ (*Indexado*) y se calculan algunas propiedades del pivote, tales como el número de elementos, el minimum bounding rectangle *MBR*, $U(\mathcal{P}_i^{\mathbb{P}})$ y $L(\mathcal{P}_i^{\mathbb{P}})$.

Utilizando el particionado basado en Diagramas de Voronoi, el algoritmo $k$CPQ [9] en MapReduce puede mejorarse modificando el cálculo local de $\beta$ y la función de *filtrado*. Para el primero, las particiones más apropiadas, aquellas donde el $k$CPQ inicial es ejecutado, son aquellas cuyos pivotes están más cerca y que tienen una mayor densidad de puntos y área de intersección. En cada partición de esta técnica, tenemos tanto su *MBR* como sus valores para $U(\mathcal{P}_i^{\mathbb{P}})$ y $L(\mathcal{P}_i^{\mathbb{P}})$, podemos detectar áreas de la partición en las que no hay puntos. Además, para la función de *filtrado* se puede utilizar una nueva distancia métrica, la mínima distancia entre pivotes (*mindist_pivots*) definida como la distancia entre pivotes menos sus valores $U(\mathcal{P}_i^{\mathbb{P}})$. De este modo, esta función elimina aquellos pares de particiones que tienen una máxima-mínima distancia ($minmaxdist = max\{mindist\_mbrs, mindist\_pivots\}$) mayor que $\beta$.

Podemos utilizar el particionado basado en Diagramas de Voronoi para mejorar la consulta $k$NNJQ en MapReduce de dos maneras: (a) realizando el *particionado inicial* de los conjuntos de datos, y/o (b) subdividiendo las particiones de $\mathbb{Q}$ en la *fase de reparticionado* y entonces utilizar sus propiedades en la fase de *kNNJ en las celdas solapadas*. Con la primera, podemos aprovechar las características de la técnica de forma global, usando los parámetros proporcionados por SpatialHadoop, de la misma manera que se hace para cualquier otra de sus consultas. Con la segunda, podemos acelerar el procesado del algoritmo $k$NNJQ descomponiendo el particionado inicial, que puede utilizar cualquier otra técnica de particionado, en particiones más pequeñas para eliminar problemas de sesgado de los datos y reducir el número y tamaño de tareas de las fases de *Bin kNNJ* y *kNNJ en las celdas solapadas*. Además, cuando se calculan las celdas que se solapan, las coordenadas de cada pivote $r_i$ y los valores de $U(\mathcal{P}_i^{\mathbb{P}})$ u $L(\mathcal{P}_i^{\mathbb{P}})$ se utilizan para mejorar el rendimiento y precisión del cálculo. Es decir, podemos detectar zonas que no contienen puntos dentro del *MBR* de una partición $\mathcal{P}_i^{\mathbb{P}}$ y por lo tanto, descartarla al no existir solape con la distancia del $k$-ésimo vecino más cercano de $p_i$.

Para finalizar, las principales conclusiones obtenidas de la realización de una serie de experimentos sobre la técnica de particionado basada en Diagramas de

Voronoi propuesta son: (1) los tiempos de particionado para $Voronoi_R$ son los más pequeños y crecen linealmente respecto al tamaño de los conjuntos de datos, mientras, para $Voronoi_k$, este incremento es mucho más grande debido a que la selección de pivotes se basa en un algoritmo *k-means*. (2) *Quadtree* mejora a todas las otras técnicas respecto al tiempo de ejecución para el algoritmo $k$CPQ, aunque ambas técnicas basadas en *Voronoi* presentan un rendimiento ligeramente inferior, especialmente, para la combinación de conjuntos de datos más pequeños. (3) Para $k$NNJQ ambas técnicas de particionado basadas en Diagramas de Voronoi son más rápidas que *Quadtree*, porque tratan mejor los datos sesgados y obtienen más resultados finales en la fase de *Bin kNNJ*. Y (4), ya que la fase de *Reparticionado* es un trabajo MapReduce, merece la pena utilizar $Voronoi_k$ en lugar de $Voronoi_R$, ya que la perdida de tiempo en la selección de pivotes con *k-means* se compensa con la ganancia de rendimiento en fases posteriores.

### 2.3 MapReduce Algorithms for the K Group Nearest-Neighbor Query

Este trabajo, que se encuentra pendiente de publicación en SAC'2019, es una colaboración internacional con la Universidad de Thessaly (Grecia) y su principal contribución es el diseño, implementación y experimentación de diferentes algoritmos MapReduce para la consulta de los ($k$) vecinos más cercanos de un grupo ((K) Group Nearest Neighbor Query (GNNQ)).

GNNQ [14] es una extensión de la consulta $k$NN que es importante para diferentes aplicaciones. Dados dos conjuntos de puntos, esta consulta obtiene los $k$ puntos de un conjunto de datos (Entrenamiento) con la menor suma de distancias a cada uno de los puntos del otro conjunto de datos (Consulta). El conjunto de Entrenamiento se considera estático y es consultado por múltiples conjuntos de datos de Consulta. Un ejemplo de su utilidad puede ser cuando tenemos un conjunto de lugares de encuentro (Entrenamiento) y un conjunto de localizaciones de usuarios (Consulta), y queremos encontrar un lugar (o $k$) de encuentro(s) que minimicen la distancia que tengan que recorrer los usuarios para llegar.

Las fases que componen el algoritmo propuesto son los siguientes:

- **Fase Preliminar**. Cálculo local del índice, lista ordenada de los puntos de Consulta, MBR, centroide y suma de distancias al centroide de los puntos de Consulta.
- **Fase 1**. Computación distribuida del número de puntos del conjunto de Entrenamiento por celda.
- **Fase 1.5**. Búsqueda local de aquellas celdas que interseccionan con el MBR de la Consulta y que contienen al menos $k$ puntos en total. El MBR se expande si es necesario.
- **Fase 2**. Computación distribuida de listas GNN, una por cada celda de la fase anterior. Se aplican heurísticas para evitar cálculos no necesarios.
- **Fase 2.5**. Combinación local de las listas GNN en una sola con los mejores puntos hasta el momento.

- **Fase 3**. Computación distribuida de listas GNN para las celdas que no intersecionan con el MBR del conjunto de Entrenamiento. Se aplican heurísticas para evitar cálculos no necesarios.
- **Fase 3.5**. Fase local (final) que combina los resultados de la Fase 2.5 con la Fase 3 en una lista GNN final.

El algoritmo propuesto utiliza 2 técnicas de particionado, Grid y Quadtree, y 2 algoritmos de cálculo, fuerza bruta y barrido del plano. Además, utiliza heurísticas y técnicas de la literatura que han sido usadas solo en sistemas centralizados. Algunas de [14, 15] se utilizan en las fases 2 y 3 para la eliminación de celdas y puntos del conjunto de Entrenamiento que no forman parte de la solución final.

Por último, tras realizar una serie de experimento sobre datos reales, se han obtenido las siguientes conclusiones: (1) El algoritmo de cálculo que mejor rendimiento da es el basado en fuerza bruta debido a que en la Fase 2 las distancias relativa son muy pequeñas y el basado en el plano no funciona bien. (2) El particionado Quadtree funciona generalmente bien, mientras el número de celdas es bajo, excepto para uno de los conjuntos de datos en el que el particionado Grid muestra una mejora de un 20 %. (3) Las heurísticas aplicadas consiguen descartar muchas celdas y puntos intermedios, evitando muchos cálculos innecesarios. (4) Se muestra mejor rendimiento al usar particiones más pequeñas ya que se puede aprovechar la computación distribuida de forma más eficiente. Finalmente (5), El MBR de la Consulta no tiene mucho impacto en los resultados, comparado con la cardinalidad y la posición relativa de ambos conjuntos.

### 2.4 Reverse K-Nearest Neighbor (R$k$NN) en MapReduce mediante el uso de técnicas de poda de particiones (SLICE)

En [7] hemos propuesto algoritmos para R$k$NN en SpatialHadoop and Location-Spark, los primeros en la literatura, para realizar de forma eficiente, paralela y distribuida el cálculo de R$k$NN sobre conjuntos reales de grandes datos espaciales. Estamos trabajando en el desarrollo de mejoras que puedan reducir el número de candidatos y por lo tanto aumentar el rendimiento.

Dado un conjunto de datos, la consulta R$k$NN [10] devuelve los objetos que tienen a un objeto de consulta dentro de sus $k$ vecinos más cercanos. Es el problema complementario a la consulta $k$NN y busca encontrar la influencia de un objeto de consulta sobre todo el conjunto de datos. En [10] se mencionan diferentes casos reales. Una solución inmediata al problema del R$k$NN es de $O(n^2)$, debido a que hay que encontrar los $k$ vecinos más cercanos de todos los $n$ objetos del conjunto de datos [10]. Obviamente, se necesitan algoritmo más eficientes y por lo tanto, la consulta R$k$NN se ha estudiado de forma extensa para entornos centralizados [19]. Nuestra propuesta trata de solucionar el problema en entornos distribuidos y busca adaptar técnicas utilizadas en estos entornos centralizados para reducir el espacio de búsqueda y mejorar el rendimiento.

Uno de estos algoritmos R$k$NN, denominado SLICE [20], esta basado en la poda basada en regiones [16] y mejora significativamente al algoritmo de referencia en términos de tiempo de ejecución. Las técnicas de poda más utilizadas son

las basadas en regiones y las de espacio medio [18]. Esta última es generalmente considerada como superior y debido a esto, casi todos los algoritmos R$k$NN utilizan y mejoran esta estrategia. Tras observar los puntos fuertes y débiles de ambas técnicas, los autores consideraron que la poda basada en regiones tiene ciertas ventajas que no habían sido aprovechadas anteriormente. El resultado de dicho estudio es la creación del algoritmo SLICE que consigue mejorar el rendimiento de este tipo de poda y eliminar sus limitaciones. Este algoritmo consta de dos fases: (1) en la fase de poda, se reduce el espacio de búsqueda e se identifica aquellos elementos *significantes* que se utilizaran para acelerar la siguiente fase. Y (2) en la fase de verificación se identifican aquellos objetos que se encuentran en el espacio de búsqueda actual y se verifican como resultado de R$k$NN si como mucho existen $k$-1 elementos significantes más cercanos al objeto de consulta.

Nuestro trabajo en curso consiste en aplicar esta filosofía en un entorno distribuido. Las fases de las que consta el algoritmo son las siguientes:

- **Fase 1**. Se realiza la fase de poda sobre la partición en la que se encuentra el objeto de consulta.
- **Fase 1.b (opcional)**. Se continua el proceso de poda sobre aquellas particiones que todavía forma parte del espacio de búsqueda.
- **Fase 2**. Se realiza la fase de verificación con aquellas particiones que no han sido eliminadas como resultado de aplicar las fases 1 y 1.b.

## 3 Conclusiones y Trabajos Futuros

El procesamiento de consultas espaciales se ha estudiado activamente en entornos centralizados, sin embargo, para marcos paralelos y distribuidos no ha conseguido una atención similar. Por lo tanto, el principal objetivo de esta tesis doctoral es, a partir del estudio y uso de diferentes entornos paralelos y distribuidos, desarrollar nuevos algoritmos eficientes que aprovechen las características que proporcionan dichos entornos. Durante el último año hemos trabajado principalmente en la mejora de algoritmos espaciales anteriormente presentados, el estudio y la implementación de nuevas consultas, el desarrollo de nuevas técnicas de particionado y la comparación de diferentes sistemas. Podemos destacar las siguientes contribuciones:

- Hemos implementado en nuevos algoritmos ($k$NNJQ y $\varepsilon$DRJQ), mejorado su rendimiento con el uso de técnicas de reparticionado de las zonas espaciales más densas y realizado la extensión de estos y de los presentados en [9, 6] para el uso de objetos espaciales complejos (polígonos).
- Hemos propuesto una técnica de particionado eficiente basada en Diagramas de Voronoi en SpatialHadoop [8] y la hemos aplicado para la mejora de los algoritmos $k$CPQ y $k$NNJQ en MapReduce.
- Hemos colaborado en el diseño, implementación y experimentación de diferentes algoritmos MapReduce para la consulta GNNQ.

– Diferentes estudios experimentales [8] de los algoritmos propuestos con conjuntos de datos sintéticos y reales han demostrado la eficiencia y escalabilidad de estos. Además se han utilizado para comparar el rendimiento de los dos DSDMS (SpatialHadoop y LocationSpark) y se ha demostrado que LocationSpark es el ganador general en cuanto a tiempo de ejecución, debido a la eficiencia del procesamiento en memoria proporcionado por Spark. Sin embargo, hay que tener en cuenta que SpatialHadoop es un DSDMS más maduro y robusto debido al tiempo dedicado a investigarlo y desarrollarlo (varios años) y proporciona más operaciones espaciales y técnicas de particionado espacial.

Trabajos futuros podrían incluir las siguiente tareas:

– Mejorar el rendimiento de $k$NNJQ mediante el uso de características de Spatial Hadoop o LocationSpark, como puede ser los *CombineFileSplits*, que permitirían reducir el tamaño de los datos enviados entre fases y la simplificación del algoritmo para la reducción de los tiempos de ejecución.
– Estudiar otros DSDMSs como GeoSpark [21] para ver si se pueden aprovechar las características que estos presentan para la mejora de las consultas ya implementadas.
– Implementar otras consultas espaciales en DSDMS, como multi-way distance joins queries [3] o basadas en la consulta R$k$NN join [5].
– Aprovechar más propiedades del particionado basado en Diagramas de Voronoi, de forma similar a como se utilizan en [12] y la comparación de nuestra propuesta con otros algoritmos MapReduce.
– Continuar con la mejora de nuestra propuesta actual de R$k$NNQ [7] mediante la adaptación y mejora de SLICE [20] para entornos distribuidos.
– Implementar otras técnicas de particionado y poda [1], porque es un factor fundamental en el rendimiento de las consultas espaciales.

## Referencias

1. Aji, A., Vo, H., Wang, F.: Effective spatial data partitioning for scalable query processing. CoRR **abs/1509.00910** (2015)
2. Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., Saltz, J.: Hadoop gis: a high performance spatial data warehousing system over mapreduce. Proceedings of the VLDB Endowment **6**(11), 1009–1020 (2013)
3. Corral, A., Manolopoulos, Y., Theodoridis, Y., Vassilakopoulos, M.: Algorithms for processing k-closest-pair queries in spatial databases. Data Knowl. Eng. **49**(1), 67–104 (2004)
4. Eldawy, A., Mokbel, M.F.: Spatialhadoop: A mapreduce framework for spatial data. In: ICDE Conference. pp. 1352–1363 (April 2015)
5. Emrich, T., Kriegel, H., Kröger, P., Niedermayer, J., Renz, M., Züfle, A.: On reverse-k-nearest-neighbor joins. GeoInformatica **19**(2), 299–330 (2015)
6. García-García, F., Corral, A., Iribarne, L., Mavrommatis, G., Vassilakopoulos, M.: A comparison of distributed spatial data management systems for processing distance join queries. In: ADBIS Conference. pp. 214–228 (2017)

7. García-García, F., Corral, A., Iribarne, L., Vassilakopoulos, M.: Rknn query processing in distributed spatial infrastructures: A performance study. In: International Conference on Model and Data Engineering. pp. 200–207. Springer (2017)
8. García-García, F., Corral, A., Iribarne, L., Vassilakopoulos, M.: Voronoi-diagram based partitioning for distance join query processing in spatialhadoop. In: International Conference on Model and Data Engineering. pp. 251–267. Springer (2018)
9. García-García, F., Corral, A., Iribarne, L., Vassilakopoulos, M., Manolopoulos, Y.: Efficient large-scale distance-based join queries in spatialhadoop. GeoInformatica (2017) https://doi.org/10.1007/s10707-017-0309-y (2017)
10. Korn, F., Muthukrishnan, S.: Influence sets based on reverse nearest neighbor queries. In: SIGMOD Conference. pp. 201–212 (May 2000)
11. Li, F., Ooi, B.C., Özsu, M.T., Wu, S.: Distributed data management using mapreduce. ACM Comput. Surv. **46**(3), 31:1–31:42 (2014)
12. Lu, W., Shen, Y., Chen, S., Ooi, B.C.: Efficient processing of k nearest neighbor joins using MapReduce. PVLDB **5**(10), 1016–1027 (2012)
13. Nodarakis, N., Pitoura, E., Sioutas, S., Tsakalidis, A.K., Tsoumakos, D., Tzimas, G.: kdann+: A rapid aknn classifier for big data. Trans. Large-Scale Data- and Knowledge-Centered Systems **24**, 139–168 (2016)
14. Papadias, D., Shen, Q., Tao, Y., Mouratidis, K.: Group nearest neighbor queries. In: Data Engineering, 2004. Proceedings. 20th International Conference on. pp. 301–312. IEEE (2004)
15. Roumelis, G., Vassilakopoulos, M., Corral, A., Manolopoulos, Y.: The k group nearest-neighbor query on non-indexed ram-resident data. In: Geographical Information Systems Theory, Applications and Management, pp. 69–89. Springer (2016)
16. Stanoi, I., Agrawal, D., El Abbadi, A.: Reverse nearest neighbor queries for dynamic databases. In: ACM SIGMOD workshop on research issues in data mining and knowledge discovery. pp. 44–53 (2000)
17. Tang, M., Yu, Y., Malluhi, Q.M., Ouzzani, M., Aref, W.G.: Locationspark: a distributed in-memory data management system for big spatial data. Proceedings of the VLDB Endowment **9**(13), 1565–1568 (2016)
18. Tao, Y., Papadias, D., Lian, X.: Reverse knn search in arbitrary dimensionality. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. pp. 744–755. VLDB Endowment (2004)
19. Yang, S., Cheema, M.A., Lin, X., Wang, W.: Reverse k nearest neighbors query processing: Experiments and analysis. PVLDB **8**(5), 605–616 (2015)
20. Yang, S., Cheema, M.A., Lin, X., Zhang, Y.: Slice: reviving regions-based pruning for reverse k nearest neighbors queries. In: Data Engineering (ICDE), 2014 IEEE 30th International Conference on. pp. 760–771. IEEE (2014)
21. Yu, J., Wu, J., Sarwat, M.: Geospark: a cluster computing framework for processing large-scale spatial data. In: SIGSPATIAL Conference. pp. 70:1–70:4 (November 2015)
22. Zhang, H., Chen, G., Ooi, B.C., Tan, K.L., Zhang, M.: In-memory big data management and processing: A survey. IEEE Transactions on Knowledge and Data Engineering **27**(7), 1920–1948 (2015)
23. Zhang, S., Han, J., Liu, Z., Wang, K., Xu, Z.: SJMR: parallelizing spatial join with MapReduce on clusters. In: CLUSTER Conference. pp. 1–8 (September 2009)

# Aportaciones desde el punto de vista del modelado y del control automático a la tecnología de destilación por membranas alimentadas con energía solar

J. D. Gil

[1] Centro Mixto CIESOL, ceiA3, Universidad de Almería.
{juandiego.gil,beren}@ual.es
[2] Centro Mixto CIESOL, CIEMAT-Plataforma Solar de Almería.
{lidia.roca}@psa.es

**Abstract.** La destilación por membranas alimentadas con energía solar (*Solar Membrane Distillation*, SMD) es una tecnología de desalación en fase de investigación, adecuada para el desarrollo de plantas autónomas capaces de cubrir requerimientos medios de demanda de agua. Esta tesis tiene como objetivo la aportación de contribuciones, desde el punto de vista de modelado control y optimización, que supongan un avance hacia la comercialización de la tecnología de destilación por membranas. En este trabajo, se presenta una revisión de los avances realizados en el año 2018 en el marco de la presente tesis, entre los cuales se puede destacar: i) la realización de un modelo basado en redes neuronales del módulo de destilación por mebranas (*Membrane Distillation,* MD), ii) el diseño y pruebas en simulación de una estrategia de arranque para la planta, basada en un controlador multivariable, y iii) el diseño y prueba en simulación de un algoritmo de control híbrido para la operación de la planta, basado en el algoritmo de control predictivo *Practical Nonlinear Predictive Control* (PNMPC).

## 1. Introducción

En las últimas décadas, la escasez de agua se está convirtiendo en uno de los principales retos que debe afrontar la humanidad. El crecimiento de la población junto con el incremento de las actividades agrícolas e industriales, han contribuido a una sobreexplotación de las reservas de agua dulce, sobrepasando el límite de renovación de dicho recurso. De este modo, las tecnologías de desalación se están convirtiendo progresivamente en un elemento necesario fundamental, especialmente en las zonas áridas o semiáridas con escasez de agua. Estas tecnologías

**Figura 1.** Instalación piloto de la PSA.

requieren intensivos sistemas de generación de energía para su funcionamiento, por lo que deben ser asociadas con fuentes de energía renovables para su sostenibilidad económica. El uso de fuentes de energía renovables en los procesos de desalación no solo reduce los costes económicos de dichos procesos, sino que también reemplaza el uso de fuentes tradicionales como los combustibles fósiles, contribuyendo así a un desarrollo medioambiental sostenible y eficiente.

En este contexto, la destilación por membranas con apoyo de energía solar es una tecnología apropiada para el desarrollo de pequeñas plantas autosuficientes de desalación, que pueden ser implantadas en zonas aisladas con unas buenas condiciones de irradiancia solar [1]. Esta tecnología destaca por su baja temperatura de operación, que permite que sea fácilmente acoplable a tecnologías solares. Además, tiene una serie de características que hacen que la planta pueda ser completamente automatizada, como son la sencillez del proceso, la fiabilidad y los bajos requerimientos de mantenimiento. No obstante, esta tecnología se encuentra actualmente en fase de investigación y todavía no ha sido comercializada a escala industrial, debido principalmente a problemas técnicos en el diseño del módulo, problemas de humectación en las membranas, a la baja producción de destilado y a la incertidumbre asociada a los costes económicos [2]. Una de las pocas plantas (ver Fig. 1) descrita totalmente en la literatura [1] se encuentra en la Plataforma Solar de Almería (PSA, www.psa.es).

MD es un proceso de separación impulsado térmicamente que se produce en una membrana hidrófoba y microporosa. La fuerza impulsora del proceso es el gradiente de presión que se genera a ambos lados de la membrana, como resultado de una diferencia de temperatura. De esta forma, las moléculas de agua se evaporan y pasan a través de la membrana, mientras que los componentes no volátiles son rechazados. Como todas las tecnologías de destilación térmica, los procesos MD tienen la capacidad de tratar agua con concentraciones de sal elevadas, pero sin necesidad de laboriosos procesos de pretratamiento químico del agua de alimentación, obteniendo permeados de alta pureza. Los sistemas MD suelen ser clasificados dependiendo del lugar donde se produzca la condensación del permeado [3]. Las configuraciones más empleadas son: Destilación por Contacto Directo (*Direct Contact MD*, DCMD) y Destilación por Membranas con Espacio de Aire (*Air Gap MD*, AGMD) en los cuales el proceso de condensación se produce dentro del módulo, y Destilación por Membranas con Barrido de Gas

(*Sweeping Gas MD*, SGMD), Destilación por Membrana Líquida (*Permeate Gap MD*, PGMD) y Destilación en Condiciones de Vacío (*Vacuum MD*, VMD) en los que la condensación tiene lugar en un condensador externo al módulo.

Como se ha mencionado anteriormente, en la PSA se encuentra una de las pocas instalaciones piloto MD que existen en la actualidad, en la cual se llevan a cabo investigaciones dirigidas a la caracterización y evaluación de diferentes tipos de módulos y configuraciones [4]. En la planta MD de la PSA (ver Fig. 2), la energía térmica necesaria para el proceso de destilación la proporciona un campo solar formado por captadores planos dispuestos en dos filas de cinco captadores cada una. La potencia nominal del campo es de 7 kW a unos 90 °C. A la salida del campo solar hay instalado un aerotermo, que se utiliza para evitar excesos de temperatura que puedan dañar las membranas. Posteriormente, se dispone un tanque de almacenamiento térmicamente aislado (1500 L), que se emplea como buffer o almacenamiento energético y que dota a la instalación de un carácter híbrido, permitiendo la producción de destilado en varios modos de operación tal y como se presentó en [5]. A continuación, se encuentra el sistema de distribución, el cual posibilita la conexión de varios módulos simultáneamente al sistema de generación de energía. Cada módulo se conecta al sistema de distribución a través de su propio intercambiador.

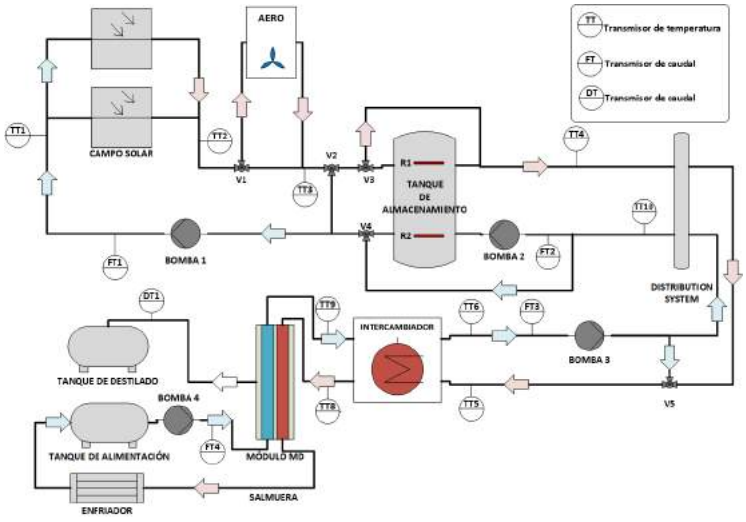



**Figura 2.** Diagrama esquemático de la instalación.

Dentro del módulo (ver Fig. 3), la bomba 4 impulsa la solución de alimentación (agua de mar) hacia el canal de condensación del módulo. Cuando dicha solución llega al intercambiador de calor, es calentada con el fluido proveniente del sistema de generación de energía. A continuación, la solución de alimentación caliente es dirigida hacia el canal de evaporación del módulo, de modo que las

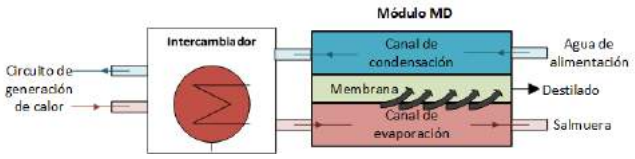



**Figura 3.** Diagrama esquemático del módulo MD.

moléculas volátiles de la solución se evaporan y pasan a través de la membrana, mientras que los componentes no volátiles se rechazan en forma de salmuera y son reconducidos al tanque de alimentación. El permeado es posteriormente condensado y vertido en el tanque de destilado. Se debe tener en cuenta que durante una operación, en el tanque de alimentación se va incrementando la temperatura y la salinidad de la solución inicial, debido a la recirculación de la salmuera. Para mantener las condiciones deseadas se utiliza un tanque auxiliar y un enfriador. Toda la planta está completamente monitorizada por medio de un PLC (*Programmable Logic Controller*) y un sistema de supervisión SCADA (*Supervisory Control And Data Acquisition*) con un tiempo de muestreo de 1 segundo.

En la literatura, hay muy pocos trabajos relacionados con el modelado, control y optimización de procesos MD. La construcción de modelos precisos de los módulos MD permite no solo simular y analizar el comportamiento del módulo bajo diferentes condiciones de operación, sino que también pueden ser usados para el desarrollo de estrategias de optimización en tiempo real, o para realizar diseños óptimos de las plantas, convirtiéndose así en elementos indispensables para la comercialización de la tecnología. Hasta el momento en la literatura, la mayoría de los modelos realizados se desarrollan en base al método estadístico RSM (*Response Surface Methodology*). Este método utiliza funciones cuadráticas para ajustar respuestas de procesos lineales o débilmente no lineales, sin embargo, no proporciona buenos resultados cuando la no linealidad del sistema es alta. Además, en la mayoría de los trabajos uno de los parámetros que más influencia el comportamiento del módulo, la salinidad del agua de alimentación (y que influye de forma no lineal al comportamiento del módulo), no se tiene en cuenta como una entrada del modelo [4,6]. Por otra parte, hay autores que utilizan modelos que se pueden ajustar de una forma más precisa a sistemas no lineales, como es el caso de las redes neuronales. Aunque en estos estudios sí se usa la salinidad como una variable de entrada al modelo, solo se considera la producción de destilado como salida [7,8,9], sin tener en cuenta el consumo térmico, uno de los factores fundamental en los procesos MD.

Desde el punto de vista del control automático, casi todos los trabajos han sido realizados en simulación, y proponen estructuras de control básicas formadas por lazos individuales. En [10], se desarrolla un modelo de una planta MD y se realiza una optimización fuera de línea de su comportamiento. Posteriormente se propone un sistema de control de temperatura basado en controladores ON/OFF que intenta seguir los puntos óptimos, obtenidos por el algoritmo

de optimización, durante una operación real simulada. En [11], nuevamente se emplean controladores del tipo ON/OFF para el control de la diferencia de temperatura que se produce entre el tanque y el campo solar, de modo que se desarrollan dos modos de operación: diario y nocturno. En [12], se prueban en simulación dos bucles de control destinados a controlar la temperatura del campo solar de alimentación. En [13], se desarrolla un generador de consignas que pretende mantener una diferencia de temperatura predefinida entre las dos partes de la membrana. Un enfoque de control más completo es el presentado en [14], en el cual se propone un algoritmo de control óptimo en tiempo real para regular el caudal de alimentación acorde a las condiciones de operación, intentando maximizar la producción de destilado. En [15], se desarrolla un modelo de red neuronal que se utiliza para el análisis del comportamiento del sistema bajo diferentes condiciones de operación, y que posteriormente es empleado para implementar un sistema de control que optimiza la producción de destilado en base a las condiciones de operación. Durante la presente tesis, ya se han presentado varias contribuciones desde el punto de vista de control, con pruebas tanto en simulación como en la planta real. En primer lugar, se desarrolló un sistema de control directo capaz de mantenerla temperatura y el caudal del sistema de generación de energía en unos valores deseados [5,16,17,18]. En segundo lugar, se han presentado trabajos donde se desarrollan algoritmos de control predictivo para la operación eficiente de la planta en tiempo real, dependiendo de las condiciones de operación en cada instante [19,20,21].

En este trabajo se presenta una revisión de los avances realizados en el marco de desarrollo de la presente tesis durante el año 2018. En particular, se ha desarrollado un modelo de predicción del módulo MD basado en datos empíricos y redes neuronales, teniendo en cuenta la salinidad del agua de alimentación como entrada al modelo. Además, se han presentado dos contribuciones desde el punto de vista de control, en el primero se desarrolló una estrategia de control multivariables para el arranque de la planta, y en el segundo, se diseñó un algoritmo de control híbrido para la operación eficiente de la planta, basado en el algoritmo de control predictivo no lineal PNMPC.

## 2. Avances

En esta sección se presentan de forma resumida los avances realizados durante este año.

### 2.1. Modelo de predicción del módulo MD

Como se ha mencionado en la sección de introducción, la obtención de modelos de predicción de los módulos MD es un factor clave para la comercialización de la tecnología. Los modelos de sistemas se pueden dividir en dos grandes grupos: i) modelos basados en primeros principios, y ii) modelos basados en datos experimentales. La construcción de un modelo basado en primeros principios es una tarea larga y laboriosa, y requiere un conocimiento total del proceso. Por el

contrario, para la realización de un modelo basado en datos experimentales no se requiere este conocimiento, sin embargo, se tiene que realizar una buena selección de las variables dependientes e independientes, así como de los experimentos a realizar. En el particular caso de los procesos MD, la dificultad para desarrollar modelos basados en primeros principios es aún mayor, debido a las diferentes configuraciones internas de los módulos, en las que la mayoría de los casos no se dispone información precisa. Por este motivo, el uso de modelos basados en datos experimentales parece más adecuado.

De este modo, y en base a los trabajos presentados hasta el momento en la literatura, se ha desarrollado un modelo de predicción basado en redes neuronales, en el cual se tienen en cuenta como entrada al modelo la temperatura a la entrada del canal de evaporación del módulo en el rango de 60-80 °C, la temperatura a la entrada del canal de condensación, 20-30 °C, el caudal de alimentación, 400-600 L/h, y la salinidad del agua de alimentación, 35-140 g/L, y como salidas la producción de destilado ($L/hm^2$) y el consumo energético ($kWh/m^3$). En la Fig. 4 se muestra la estructura de la red. El trabajo al completo fue presentado en [22]. Además, en este trabajo se realizó también la comparación del modelo de red neuronal con uno basado en RSM, analizando así las ventajas de lo modelos de red neuronal en comportamientos no lineales, como el que ocasiona la salinidad del agua de alimentación en la producción y el consumo energético del módulo MD.

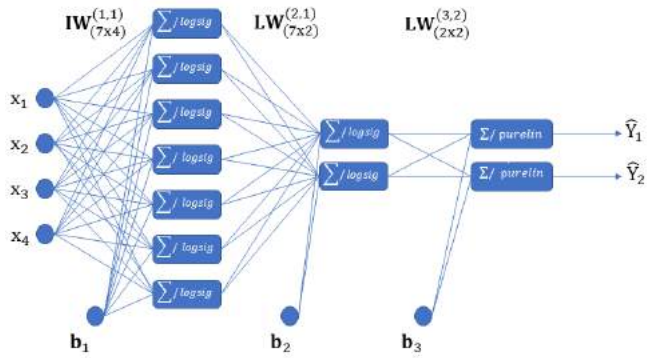



**Figura 4.** Estructura de la red neuronal (*Multi-Layer feedforward Perceptron*, MLP). Todas las variables presentes se explican con detalle en [22].

## 2.2. Algoritmo de control multivariable para el arranque de la planta

El algoritmo de control multivariable para el arranque de la planta se planteó como una mejora a la estrategia de arranque presentada en [21]. De este forma, se diseñó un algoritmo de control multivariable TITO (*Two Imputs Two Outputs*)

complementado con un generador de referencias en el que se resuelve un problema de optimización en tiempo real. Por un lado, el generador de referencias calcula las referencias óptimas para el controlador multivariable en cada instante de muestreo, usando un modelo estático del campo solar. Por otro lado, la estructura de control multivariable está formada por controladores PID combinados con controladores por adelanto y desacopladores, los cuales se encargan de seguir las referencias calculadas por el generador de referencias. Las variables controladas son la temperatura de entrada y de salida del campo solar. La temperatura de salida se controla mediante un esquema de control en cascada, mientras que la temperatura de entrada se calcula con un predictor de Smith filtrado. El diagrama esquemático del sistema de control se presenta en la Fig. 5 y el trabajo al completo se puede ver en [23].

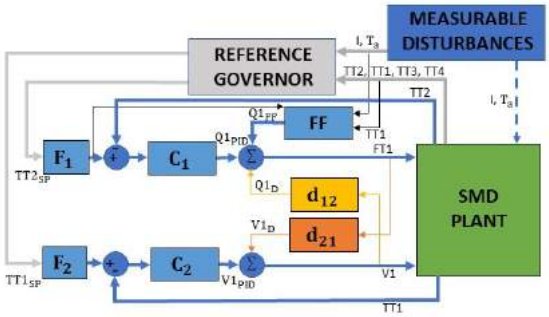



**Figura 5.** Diagrama esquemático de la estructura de control [23].

## 2.3. Algoritmo de control híbrido para la operación eficiente de la planta

A diferencia del enfoque presentado en [21], en este trabajo se tienen en cuenta los cambios entre los diferentes modos de operación de la planta en el horizonte de predicción. Para ello, se consideró la planta como un sistema MLD (*Mixed Logycal Dynamical*), y se diseñó un algoritmo hybrido basado en la metodología PNMPC, la cual proporciona la acciones de control óptimas en cada instante de muestreo resolviendo un problema de optimización MILP (*Mixed Integer Lineal Programming*). Los objetivos del sistema de control son maximizar el tiempo de operación de la planta, la temperatura de operación y la producción de destilado. De este modo, se definieron 5 modos de operación, los cuales se introdujeron en el problema de control por medio de restricciones operacionales. Así, el sistema de control selecciona el modo de operación y los puntos de operación óptimos en términos de caudal de alimentación del módulo MD acorde con las condiciones de operación. El trabajo al completo se presentó en [24].

## 3. Conclusiones

En esta sección se presentan las principales conclusiones de los avances realizados durante este año. En primer lugar, el modelo basado en redes neuronales del módulo MD presentó una mayor precisión que la técnica RSM, especialmente en el ajuste del consumo térmico del módulo. Esto se debe a que la salinidad del agua de alimentación afecta de una forma no lineal al consumo térmico del módulo MD, la cual no puede ser representada mediante una ecuación cuadrática. Por tanto, las redes neuronales parecen más adecuadas para el desarrollo de modelos precisos de este tipo de sistemas cuando se tiene en cuenta la salinidad del agua de alimentación como una entrada. Sin embargo, también se debe mencionar que este tipo de modelos requiere más datos experimentales que el RSM para su correcto desarrollo.

En lo relativo al sistema de control multivariable para el arranque de la planta, su comportamiento se probó en simulación, y se comparó con una operación manual convencional. Los resultados mostraron que la producción de destilado puede aumentar en un 6 % en comparación con dicha operación, debido principalmente a que se consigue arrancar la planta en torno a 15 minutos antes.

Por último, el algoritmo de control híbrido para la operación eficiente de la planta fue probado también en simulación. En este caso, los resultados se compararon con una operación con una máquina de estados. Esta comparación mostró que con el algoritmo propuesto, la producción de destilado puede aumentar en torno a 1.40 %, y la operación se puede extender en un 11 %.

## Referencias

1. G. Zaragoza, A. Ruiz-Aguirre, and E. Guillén-Burrieza, "Efficiency in the use of solar thermal energy of small membrane desalination systems for decentralized water production," *Applied Energy*, vol. 130, pp. 491–499, 2014.
2. M. Khayet, "Solar desalination by membrane distillation: Dispersion in energy consumption analysis and water production costs (a review)," *Desalination*, vol. 308, pp. 89–101, 2013.
3. A. Alkhudhiri, N. Darwish, and N. Hilal, "Membrane distillation: a comprehensive review," *Desalination*, vol. 287, pp. 2–18, 2012.
4. A. Ruiz-Aguirre, J. Andres-Manas, J. Fernández-Sevilla, and G. Zaragoza, "Modeling and optimization of a commercial permeate gap spiral wound membrane distillation module for seawater desalination," *Desalination*, vol. 419, pp. 160–168, 2017.
5. J. D. Gil, A. Ruiz-Aguirre, L. Roca, G. Zaragoza, and M. Berenguel, "Solar membrane distillation: A control perspective," in *23th Mediterranean Conference on Control and Automation (MED 2015)*. Torremolinos, Málaga, Spain, 2015, pp. 796–802.
6. M. Khayet, C. Cojocaru, and A. Baroudi, "Modeling and optimization of sweeping gas membrane distillation," *Desalination*, vol. 287, pp. 159–166, 2012.
7. M. Khayet, C. Cojocaru, and M. Essalhi, "Artificial neural network modeling and response surface methodology of desalination by reverse osmosis," *Journal of Membrane Science*, vol. 368, no. 1-2, pp. 202–214, 2011.

8. M. Khayet and C. Cojocaru, "Artificial neural network modeling and optimization of desalination by air gap membrane distillation," *Separation and Purification Technology*, vol. 86, pp. 171–182, 2012.
9. ——, "Artificial neural network model for desalination by sweeping gas membrane distillation," *Desalination*, vol. 308, pp. 102–110, 2013.
10. H. Chang, G.-B. Wang, Y.-H. Chen, C.-C. Li, and C.-L. Chang, "Modeling and optimization of a solar driven membrane distillation desalination system," *Renewable Energy*, vol. 35, no. 12, pp. 2714–2722, 2010.
11. H. Chang, S.-G. Lyu, C.-M. Tsai, Y.-H. Chen, T.-W. Cheng, and Y.-H. Chou, "Experimental and simulation study of a solar thermal driven membrane distillation desalination process," *Desalination*, vol. 286, pp. 400–411, 2012.
12. J.-S. Lin, H. Chang, and G. B. Wang, "Modelling and control of the solar powered membrane distillation system," in *AIChE Annual Meeting*. Minneapolis, MN, USA, 2011.
13. F. Eleiwi, I. N'Doye, and T.-M. Laleg-Kirati, "Feedback control for distributed heat transfer mechanisms in direct-contact membrane distillation system," in *2015 IEEE Conference on Control Applications (CCA)*. Sydney, Australia, 2015, pp. 1624–1629.
14. A. M. Karam and T. M. Laleg-Kirati, "Real time optimization of solar powered direct contact membrane distillation based on multivariable extremum seeking," in *Control Applications (CCA), 2015 IEEE Conference on*. IEEE, 2015, pp. 1618–1623.
15. R. Porrazzo, A. Cipollina, M. Galluzzo, and G. Micale, "A neural network-based optimizing control system for a seawater-desalination solar-powered membrane distillation unit," *Computers & Chemical Engineering*, vol. 54, pp. 79–96, 2013.
16. J. D. Gil, A. Ruiz-Aguirre, L. Roca, G. Zaragoza, M. Berenguel, and J. L. Guzmán, "Control de plantas de destilación por membranas con apoyo de energía solar–parte 1: Esquemas," in *XXXVI Jornadas Automática, Bilbao, España*, 2015.
17. ——, "Control de plantas de destilación por membranas con apoyo de energía solar–parte 2: Resultados," in *XXXVI Jornadas Automática, Bilbao, España*, 2015.
18. J. D. Gil, L. Roca, G. Zaragoza, and M. Berenguel, "A feedback control system with reference governor for a solar membrane distillation pilot facility," *Renewable Energy*, vol. 120, pp. 536–549, 2018.
19. J. D. Gil, L. Roca, M. Berenguel, A. Ruiz, G. Zaragoza, and A. Gimenez, "Control predictivo para la operación eficiente de una planta formada por un sistema de desalación solar y un invernadero," in *XXXVIII Jornadas Automática, Gijón, España*, 2017.
20. J. D. Gil, L. Roca, A. Ruiz-Aguirre, G. Zaragoza, J. L. Guzmán, and M. Berenguel, "Using a nonlinear model predictive control strategy for the efficient operation of a solar-powered membrane distillation system," in *25th Mediterranean Conference on Control and Automation (MED 2017)*. Valleta, Malta, 2017.
21. J. D. Gil, L. Roca, A. Ruiz-Aguirre, G. Zaragoza, and M. Berenguel, "Optimal operation of a solar membrane distillation pilot plant via nonlinear model predictive control," *Computers & Chemical Engineering*, vol. 109, pp. 151–165, 2018.
22. J. D. Gil, A. Ruiz-Aguirre, L. Roca, G. Zaragoza, and M. Berenguel, "Prediction models to analyse the performance of a commercial-scale membrane distillation unit for desalting brines from RO plants," *Desalination*, vol. 445, pp. 15–28, 2018.
23. J. D. Gil, L. Roca, M. Berenguel, and J. L. Guzman, "A multivariable controller for the start-up procedure of a solar membrane distillation facility," in *3rd IFAC conference on Advances in Proportional-Integral-Derivative Control, Gante, Belgica*, 2018.

24. J. D. Gil, P. Mendes, G. Andrade, L. Roca, J. Normey-Rico, and M. Berenguel, "Hybrid NMPC applied to a solar-powered membrane distillation system, In press," in *Dynamics and Control of Process Systems, including Biosystems - 12th DYCOPS, Florianópolis, Brasil*, 2019.

# Aceleración del filtro basado en Difusión No-Lineal Anisótropa

Juan José Moreno Riado[1]

Departamento de Informática, Universidad de Almería
juanjomoreno@ual.es

**Resumen** El filtro basado en Difusión No-Lineal Anisótropa (AND) es actualmente la técnica predominante para el filtrado en tomografía electrónica. Utilizando algoritmos especializados y técnicas de computación de alto rendimiento podemos optimizar la ejecución del filtro AND, reduciendo tanto el tiempo de ejecución como los requerimientos de memoria. La herramienta resultante permite filtrar grandes volúmenes rápidamente en ordenadores convencionales.

**Keywords:** AND, Tomografía electrónica, Reducción de ruido, HPC.

## 1. Introducción

La tomografía electrónica (TE) es una técnica importante para la visualización tridimensional (3D) de la arquitectura celular con resolución nanométrica, permitiendo el abordaje de problemas fundamentales en biología celular y molecular [8,2]. La TE es similar a la tomografía computerizada en Medicina, pero haciendo uso de un microscopio electrónico. A partir de una serie de imágenes de la muestra adquiridas a distintas vistas, se realizan un conjunto de procesos computacionales que culminan con la obtención de un volumen 3D [3,8].

El procesamiento de imagen es un componente fundamental en los estudios estructurales por TE [3]. En primer lugar es esencial para la obtención del volumen 3D o tomograma, que se calcula por medio de métodos de reconstrucción tomográfica que operan sobre las imágenes adquiridas. En segundo lugar, también es necesario para la interpretación del volumen 3D debido a la complejidad de la información que contiene, al nivel de ruido y a las características/limitaciones inherentes a la técnica. Así, se emplean de forma habitual etapas de reducción de ruido, segmentación o identificación de características estructurales, y finalmente análisis de subestructuras o patrones repetitivos.

Algunas de las etapas de la TE presentan unas demandas computacionales especialmente altas, derivadas tanto del volumen de datos a procesar como de la complejidad algorítmica de los métodos. Las técnicas de computación de altas prestaciones (HPC) ha jugado un papel esencial en la gestión eficiente de estos procesos en distintas plataformas [5].

La interpretación de los volúmenes suele complicarse por la baja relación señal ruido (SNR) de los tomogramas, especialmente en condiciones criogénicas.

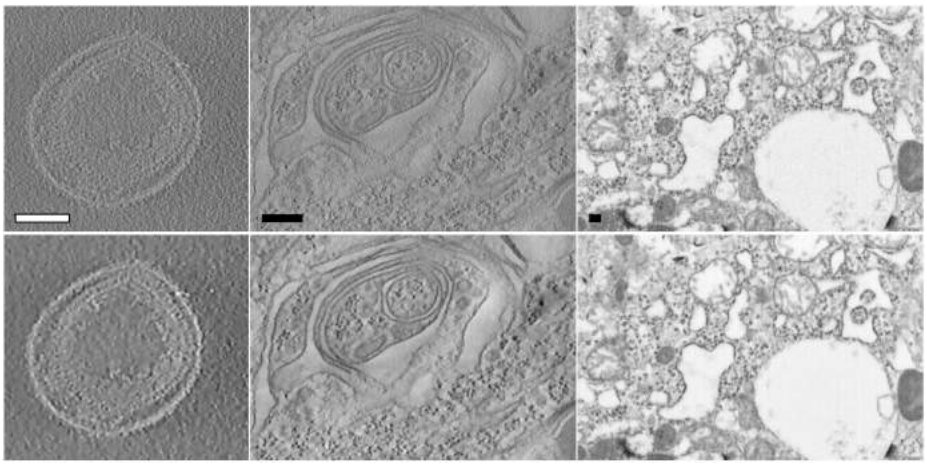


**Figura 1.** Reducción de ruido con AND. Ejemplos de su aplicación a volúmenes 3D de muestras biológicas obtenidos por distintas técnicas de microscopía electrónica. Destaca la buena preservación de las membranas en todos los casos mientras que el ruido ha sido reducido sustancialmente. Escala línea blanca: 100 nm, líneas negras: 200 nm.

Por lo tanto, es usual aplicar técnicas de reducción de ruido en la etapa de postprocesamiento [3] o durante la reconstrucción tridimensional [1]. Esta misma necesidad de filtrado aparece en otras técnicas de microscopía electrónica 3D para visualizar estructuras subcelulares [10].

Gracias a su capacidad de reducir ruido preservando los rasgos de la imagen, el filtro basado en Difusión No-Lineal Anisótropa (AND) es actualmente la técnica predominante de filtrado en ET [6,4]. La Figura 1 muestra algunos ejemplos de la aplicación de AND sobre volúmenes 3D de muestras biológicas.

## 2. Descripción de AND

AND ajusta la intensidad y la dirección del filtrado a partir de la estructura local alrededor de cada vóxel, estimada por el análisis de los autovalores y autovectores del tensor de estructura:

$$\mathbf{J}(\mathrm{I}) = \nabla \mathrm{I} \cdot \nabla \mathrm{I}^{\mathrm{T}} = \begin{bmatrix} \mathrm{I}_x^2 & \mathrm{I}_x\mathrm{I}_y & \mathrm{I}_x\mathrm{I}_z \\ \mathrm{I}_x\mathrm{I}_y & \mathrm{I}_y^2 & \mathrm{I}_y\mathrm{I}_z \\ \mathrm{I}_x\mathrm{I}_z & \mathrm{I}_y\mathrm{I}_z & \mathrm{I}_z^2 \end{bmatrix} = \mathbf{V}\mathbf{Q}\mathbf{V}^{\mathrm{T}} \tag{1}$$

donde $\nabla \mathrm{I} = (\mathrm{I}_x, \mathrm{I}_y, \mathrm{I}_z)$ es el vector gradiente del volumen I y $\mathbf{V}\mathbf{Q}\mathbf{V}^{\mathrm{T}}$ es la descomposición en valores propios de $\mathbf{J}$. AND sigue la ecuación de difusión:

$$\mathrm{I}_t = \mathrm{div}(\mathbf{D} \cdot \nabla \mathrm{I}) \tag{2}$$

donde $\mathrm{I}_t$ define la derivada con respecto al tiempo y div es el operador de divergencia. La matriz de $3 \times 3$ $\mathbf{D}$ es el tensor de difusión que ajusta el filtrado acorde

a la estructura local. $\mathbf{D}$ se construye a partir de los autovectores $\mathbf{v_i}$ del tensor de estructura (Ec. 1) y sus autovalores $\lambda_i$ (con valores en el rango $[0,1]$) definen la intensidad del suavizado en la dirección $\mathbf{v_i}$ correspondiente:

$$\mathbf{D}(\mathbf{J}) = \mathbf{V}\mathbf{L}\mathbf{V}^{\mathsf{T}} = [\mathbf{v_1}\ \mathbf{v_2}\ \mathbf{v_3}] \cdot \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \cdot [\mathbf{v_1}\ \mathbf{v_2}\ \mathbf{v_3}]^{\mathsf{T}} \tag{3}$$

Para conservar los bordes, el suavizado en la dirección de máxima variación de densidad ($\mathbf{v_1}$) se fija como una función monótona decreciente del gradiente. Normalmente, $\lambda_1 = 1{,}0 - \exp(-3{,}31488/(|\nabla \mathrm{I}|/\mathrm{K})^8)$, donde el parámetro $\mathrm{K}$ actúa como el umbral del gradiente que define a los bordes. En cambio, $\lambda_2$=$\lambda_3$=1 son definidos para filtrar drásticamente en las dos direcciones de menor variación.

La ecuación de difusión (Ec. 2) puede ser aproximada numéricamente utilizado diferencias finitas. El término $\mathrm{I_t} = \frac{\partial \mathrm{I}}{\partial \mathrm{t}}$ puede ser reemplazado por una diferencia progresiva de Euler. El esquema explícito resultante permite el cálculo de las sucesivas versiones del volumen iterativamente:

$$\begin{aligned} \mathrm{I}^{(\mathrm{k})} = \mathrm{I}^{(\mathrm{k}-1)} + \tau \cdot (\ & \tfrac{\partial}{\partial \mathrm{x}}(\mathrm{D}_{11}\mathrm{I_x}) + \tfrac{\partial}{\partial \mathrm{x}}(\mathrm{D}_{12}\mathrm{I_y}) + \tfrac{\partial}{\partial \mathrm{x}}(\mathrm{D}_{13}\mathrm{I_z}) + \\ & \tfrac{\partial}{\partial \mathrm{y}}(\mathrm{D}_{21}\mathrm{I_x}) + \tfrac{\partial}{\partial \mathrm{y}}(\mathrm{D}_{22}\mathrm{I_y}) + \tfrac{\partial}{\partial \mathrm{y}}(\mathrm{D}_{23}\mathrm{I_z}) + \\ & \tfrac{\partial}{\partial z}(\mathrm{D}_{31}\mathrm{I_x}) + \tfrac{\partial}{\partial z}(\mathrm{D}_{32}\mathrm{I_y}) + \tfrac{\partial}{\partial z}(\mathrm{D}_{33}\mathrm{I_z})) \end{aligned} \tag{4}$$

donde $\tau$ indica la longitud del paso temporal, $\mathrm{I}^{(\mathrm{k})}$ indica el volumen en el instante $\mathrm{t_k} = \mathrm{k}\tau$ y el término $\mathrm{D_{mn}}$ representa los componentes del tensor de difusión $\mathbf{D}$, con $\mathrm{D_{mn}} = \mathrm{D_{nm}}$ debido a la simetría. Las derivadas espaciales ($\frac{\partial}{\partial \mathrm{x}}$, $\frac{\partial}{\partial \mathrm{y}}$ y $\frac{\partial}{\partial z}$) son aproximadas basándose en las diferencias centrales. Para mantener estabilidad numérica, el máximo paso temporal en este caso es $\tau = 0{,}1$.

Aunque AND es una potente técnica de filtrado, es computacionalmente costosa en términos de tiempo de ejecución y consumo de memoria, lo que impide su aplicación a grandes volúmenes.

## 3. Autovectores y autovalores

La diagonalización numérica de una matriz real simétrica $\mathbf{A}$ consiste en calcular un conjunto de autovalores $\lambda_i$ y autovectores $\mathbf{v_i}$ que satisfacen $\mathbf{A}\mathbf{v_i} = \lambda_i \mathbf{v_i}$, resultando en la siguiente factorización:

$$\mathbf{A} = \mathbf{V}\mathbf{L}\mathbf{V}^{\mathsf{T}} = [\mathbf{v_1}\ \mathbf{v_2}\ \mathbf{v_3}] \cdot \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \cdot [\mathbf{v_1}\ \mathbf{v_2}\ \mathbf{v_3}]^{\mathsf{T}} \tag{5}$$

### 3.1. Algoritmo de Jacobi

El método más preciso para la diagonalización de matrices es el Algoritmo de Jacobi [11]. Consiste en aplicar una secuencia de transformaciones de similitud ortogonales de la forma $\mathbf{A} \rightarrow \mathbf{P}_j \ \mathbf{A} \ \mathbf{P}_j^T$:

$$\mathbf{P}_j = \begin{bmatrix} 1 & & & & & \\ & \cdots & & & & \\ & & c & \cdots & s & \\ & & \vdots & 1 & \vdots & \\ & & -s & \cdots & c & \\ & & & & & \cdots \\ & & & & & 1 \end{bmatrix} \tag{6}$$

Cada transformación, llamada una *rotación de Jacobi*, tiene como objetivo poner a cero (o hacer insignificante) un elemento fuera de la diagonal, convirtiendo eventualmente $\mathbf{A}$ en una matriz diagonal, dentro del épsilon de la máquina.

El producto de las transformaciones, $\prod_j \mathbf{P}_j$, nos da la matriz de autovectores mientras que los elementos de la matriz diagonal resultante son los autovalores. Normalmente, las matrices necesitan un número de transformaciones igual a $3 - 5\times$ el número de elementos (por ejemplo, entre 27 y 45 para matrices de $3 \times 3$) para converger. Cada una de estas transformaciones tienen del orden de $12 \times 3$ operaciones [11,7].

### 3.2. Cálculo analítico directo

La forma más rápida de calcular los autovalores de una matriz simétrica $\mathbf{A}$ de $3 \times 3$ es la resolución de su ecuación característica:

$$P(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}) = 0 \tag{7}$$

donde $\mathbf{I}$ es la matriz identidad y det es el determinante de la matriz. Siendo $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix} \tag{8}$$

La ecuación característica 7 puede ser expresada como una ecuación cúbica:

$$P(\lambda) = \lambda^3 + c_2\lambda^2 + c_1\lambda + c_0 = 0 \tag{9}$$

con los coeficientes:

$$\begin{aligned} c_2 &= -a_{11} - a_{22} - a_{33} \\ c_1 &= a_{11}a_{22} + a_{11}a_{33} + a_{22}a_{33} - a_{12}^2 - a_{13}^2 - a_{23}^2 \\ c_0 &= a_{11}a_{23}^2 + a_{22}a_{13}^2 + a_{33}a_{12}^2 - a_{11}a_{22}a_{33} - 2a_{13}a_{12}a_{23} \end{aligned} \tag{10}$$

Kopp [7] demuestra en su artículo que esta ecuación se puede resolver utilizando el método de Cardano y obtiene expresiones analíticas para los tres autovalores:

$$\lambda_1 = -\tfrac{1}{3}(c_2 + c) + c \qquad \lambda_2 = -\tfrac{1}{3}(c_2 + c) - s \qquad \lambda_3 = -\tfrac{1}{3}(c_2 + c) + s \tag{11}$$

donde:

$$\begin{gathered} c = \sqrt{p}\cos\phi\ , \qquad s = \tfrac{1}{\sqrt{3}}\sqrt{p}\sin\phi\ , \qquad p = c_2^2 - 3c_1 \\ \phi = \tfrac{1}{3}\arctan\frac{\sqrt{27\left[\frac{1}{4}c_1^2(p - c_1) + c_0\left(q + \frac{27}{4}c_0\right)\right]}}{q}\ , \qquad q = -c_2\left(p - \tfrac{3}{2}c_1\right) - \tfrac{27}{2}c_0 \end{gathered} \tag{12}$$

Tras esto, los autovectores pueden ser eficientemente calculados con productos vectoriales. Todos los autovectores $\mathbf{v_i}$ satisfacen, por definición:

$$(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{v}_i = 0 \tag{13}$$

Si calculamos el conjugado hermítico de esta ecuación y lo multiplicamos por un vector $\mathbf{x}$, obtenemos:

$$\mathbf{v}_i^{\mathsf{T}}(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{x} = 0 \tag{14}$$

Esto se cumple para cualquier $\mathbf{x} \in \mathbb{C}^3$ y en particular para los vectores unitarios $\mathbf{e}_1 = (1,0,0)^{\mathsf{T}}$ y $\mathbf{e}_2 = (0,1,0)^{\mathsf{T}}$. Consecuentemente, $\mathbf{v}_i$ se puede calcular como:

$$\begin{aligned} \mathbf{v}_1 &= (\mathbf{A}\mathbf{e}_1 - \lambda_1\mathbf{e}_1) \times (\mathbf{A}\mathbf{e}_2 - \lambda_1\mathbf{e}_2) \\ \mathbf{v}_2 &= (\mathbf{A}\mathbf{e}_1 - \lambda_2\mathbf{e}_1) \times (\mathbf{A}\mathbf{e}_2 - \lambda_2\mathbf{e}_2) \\ \mathbf{v}_3 &= \mathbf{v}_1 \times \mathbf{v}_2 \end{aligned} \tag{15}$$

El método analítico de Kopp es muy rápido, siendo las operaciones más costosas las funciones trigonométricas $(\sin, \cos, \arctan)$ de la Ecuación 12. Aunque este método es propenso a tener una baja precisión numérica, esta desventaja no es un problema para aplicaciones que solo requieren una precisión moderada [7]. Kopp también propone un algoritmo híbrido en el cual, dado el caso de que el cálculo analítico no sea lo suficientemente preciso, se utilice otro algoritmo de mayor precisión. En nuestra implementación hemos utilizado el método analítico puro, ya que no hemos observado diferencias significantes en los volúmenes estudiados.

## 4. Evaluación

En esta sección evaluaremos las versiones de AND que implementan los algoritmos de cálculo de autovalores y autovectores de la Sección 3. Disponemos de cuatro versiones: Dos Multicore implementadas en C con Pthreads y dos GPU implementadas en CUDA C. Como referencia utilizaremos una versión secuencial de AND que implementa un algoritmo Jacobi de propósito general, es decir, no optimizado para matrices de $3 \times 3$.

Como plataformas experimentales utilizaremos las plataformas $\mathcal{F}_1$ y $\mathcal{F}_2$ descritas en la Tabla 1. Se han seleccionado estas plataformas por ser las que mejor rendimiento proporcionan para las implementaciones CUDA y Multicore, respectivamente.

**Tabla 1.** Especificaciones técnicas de las plataformas $\mathcal{F}_1$ y $\mathcal{F}_2$.

| $\mathcal{F}_1$ | Bullx R421-E4 |
|---|---|
| **CPU** | 2 x Intel Xeon E5-2620 v3 (Haswell, 6 núcleos) @ 2.4 GHz |
| **GPU** | 2 x NVIDIA Tesla K80 (arquitectura Kepler) |
| **RAM** | 64 GB DDR4 @ 2133 MHz |
| **HDD** | 1 TB SATA3 |
| $\mathcal{F}_2$ | Bullx R424-E3 |
| **CPU** | 2 x Intel Xeon E5-2650 v2 (Ivy Bridge, 8 núcleos) @ 2.6 GHz |
| **RAM** | 128 GB DDR3 @ 1866 MHz |
| **HDD** | 1 TB SATA3 |

### 4.1. Rendimiento y consumo energético

Comenzamos la evaluación con un análisis del rendimiento y el consumo energético de las implementaciones paralelas. Para este análisis utilizaremos versiones reescaladas de un mismo volumen. El volumen original es un cubo, es decir, tiene el mismo número de vóxeles en cada una de sus tres dimensiones. Como los volúmenes reescalados tienen esa misma geometría, utilizaremos el número de vóxels por dimensión para identificarlos. De esta forma, tenemos cinco volúmenes de prueba: 256, 384, 460, 512 y 640. Para cada volumen, implementación y plataforma se ha realizado 100 experimentos, descartando los 10 mejores y los 10 peores y haciendo la media de los restantes. En cada uno de los experimentos se han aplicado 10 iteraciones de AND, que suele ser lo común en el campo de la ET.

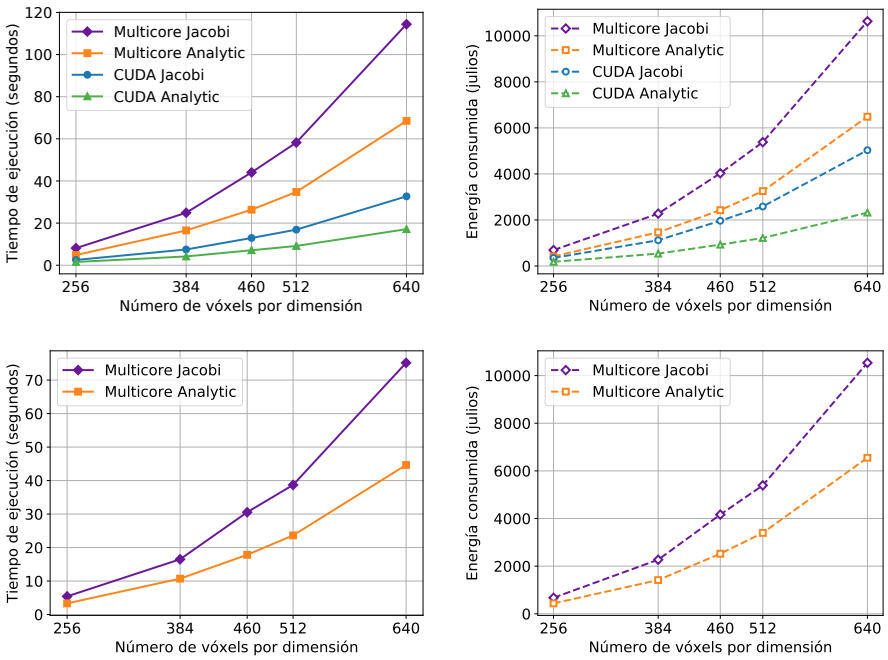



**Figura 2.** Tiempo de ejecución (izquierda, líneas continuas) y Energía consumida (derecha, líneas discontinuas) de las cuatro implementaciones paralelas de AND en la plataforma $\mathcal{F}_1$ (arriba) y la plataforma $\mathcal{F}_2$ (abajo) para los cinco volúmenes de prueba.

Para las dos plataformas, la Figura 2 muestra, a la izquierda, los tiempos de ejecución en segundos y, a la derecha, una estimación de la energía consumida en julios por cada implementación. Podemos observar que las implementaciones analíticas proporcionan mejor rendimiento y menor consumo energético que las implementaciones Jacobi, tanto en las GPUs como en los Multicore. También se observa que las versiones GPU son más rápidas y consumen menos energía que las Multicore. Esto se debe a que, aunque AND es un algoritmo iterativo con un conjunto de pasos serializados, de cada uno de esos pasos se puede extraer, internamente, un gran paralelismo.

A continuación, evaluaremos la aceleración y la eficiencia energética de los algoritmos paralelos. Como referencia, utilizaremos una implementación secuencial cuyos tiempos de ejecución y consumos energéticos son descritos en la Tabla 2 para las plataformas $\mathcal{F}_1$ y $\mathcal{F}_2$.

La Figura 3 muestra la aceleración y la eficiencia energética para las dos plataformas. La principal información que se puede extraer de estas figuras es la pequeña pendiente de las líneas, de la que podemos concluir que las implementaciones paralelas escalan adecuadamente para todos los tamaños estudiados.

**Tabla 2.** Tiempos de ejecución y consumos energéticos de la implementación secuencial de referencia para los cinco volúmenes de prueba.

| | **Plataforma $\mathcal{F}_1$** | | **Plataforma $\mathcal{F}_2$** | |
|---|---|---|---|---|
| **Volumen** | Tiempo (seg.) | Energía (J) | Tiempo (seg.) | Energía (J) |
| $256^3$ | 65.68 | 2214.42 | 65.74 | 3254.46 |
| $384^3$ | 223.81 | 7611.79 | 226.05 | 11200.32 |
| $460^3$ | 381.30 | 12988.92 | 384.49 | 19098.48 |
| $512^3$ | 544.12 | 18493.36 | 542.18 | 26948.10 |
| $640^3$ | 1086.07 | 36862.97 | 1071.25 | 53238.07 |

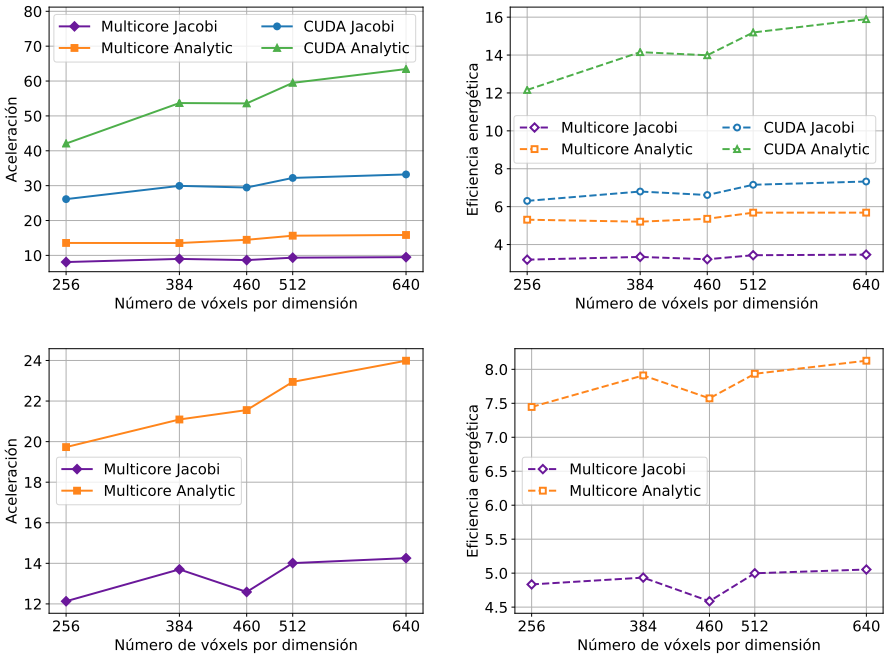



**Figura 3.** Aceleración (izquierda, líneas continuas) y Eficiencia energética (derecha, líneas discontinuas) de las cuatro implementaciones paralelas de AND en la plataforma $\mathcal{F}_1$ (arriba) y la plataforma $\mathcal{F}_2$ (abajo) para los cinco volúmenes de prueba.

En la plataforma $\mathcal{F}_1$, el factor de aceleración del algoritmo analítico frente al Jacobi es $\sim 1{,}76\times$ para las versiones GPU y $\sim 1{,}61\times$ para las versiones Multicore. El factor de eficiencia energética es $\sim 2{,}05\times$ para las GPU y $\sim 1{,}63\times$ para las Multicore. En la plataforma $\mathcal{F}_2$, el factor de aceleración es $\sim 1{,}68\times$ y el de eficiencia energética $\sim 1{,}59\times$ para la versión Multicore.

Igual que anteriormente, las implementaciones analíticas consiguen mayores aceleraciones y mejores eficiencias energéticas que las equivalentes Jacobi. Para las implementaciones Multicore, en la plataforma $\mathcal{F}_2$ se obtienen mejores métricas que en la plataforma $\mathcal{F}_1$, debido al mayor número de núcleos por procesador.

### 4.2. Precisión del algoritmo analítico

Aunque la implementación analítica es notablemente más rápida que la Jacobi, en estos filtros además de la rapidez, la calidad del filtrado es muy importante. En esta sección presentamos métricas para juzgar si el resultado obtenido con el algoritmo analítico es adecuado, en comparación con el algoritmo Jacobi.

La primera métrica es el error relativo entre los volúmenes resultantes de los dos algoritmos. Se puede formular como:

$$\frac{\sum_{i=1}^{N} |I_i^j - I_i^a|}{\sum_{i=1}^{N} I_i^j} \tag{16}$$

donde $N$ es el número de vóxels del volumen y $I_i^j$ y $I_i^a$ son los vóxels de la solución obtenida con los algoritmos Jacobi y analítico, respectivamente.

Aunque esta métrica es adecuada, su carácter global puede esconder vóxels concretos en los que la diferencia sea grande. Por lo tanto, también conviene calcular el error relativo máximo de todos los pares de vóxels:

$$\max_{i=1\ldots N} \{\frac{|I_i^j - I_i^a|}{\frac{1}{N}\sum_{i=1}^{N} I_i^j}\} \tag{17}$$

La tabla 3 muestra estas dos métricas para los volúmenes anteriormente estudiados. Se puede observar que el error relativo es insignificante y que el máximo error relativo individual en los pares de vóxels está por debajo del 1 %.

**Tabla 3.** Error relativo y máximo error relativo individual entre volúmenes filtrados utilizando el algoritmo Jacobi y el analítico.

| Volumen | Error relativo | Máx. error rel. ind. |
|---|---|---|
| $256^3$ | 0.00000655 | 0.00718860 |
| $384^3$ | 0.00000757 | 0.00754244 |
| $460^3$ | 0.00000756 | 0.00363749 |
| $512^3$ | 0.00000779 | 0.00780323 |
| $640^3$ | 0.00000768 | 0.00585213 |

### 4.3. Conclusiones

Este trabajo ha dado como resultado TomoEED, una herramienta software basada en el filtro AND para la reducción de ruido preservando los rasgos de los tomogramas. Ambas implementaciones paralelas mejoran el rendimiento del filtro en factores que alcanzan 62× y reducen el consumo de energía en factores de hasta 16×. También se confirma que el algoritmo analítico proporciona un gran aumento de velocidad frente al tradicional Jacobi y se demuestra que no causa una pérdida significativa de la calidad de la solución.

El software se ha puesto a disposición de la comunidad científica [1] y se ha publicado un artículo en la revista *Bioinformatics* difundiendo el trabajo realizado [9].

## Referencias

1. Chen, Y., et al.: FIRT: Filtered iterative reconstruction technique with information restoration. J. Struct. Biol. **195**, 49–61 (2016)
2. Doerr, A.: Cryo-electron tomography. Nature Methods **14**(1), 34 (2016)
3. Fernandez, J.J.: Computational methods for electron tomography. Micron **43**, 1010–1030 (2012)
4. Fernandez, J.J., Li, S.: An improved algorithm for anisotropic diffusion for denoising tomograms. J. Struct. Biol. **144**, 152–161 (2003)
5. Fernandez, J.J.: High performance computing in structural determination by electron cryomicroscopy. Journal of structural biology **164**(1), 1–6 (2008)
6. Frangakis, A.S., Hegerl, R.: Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. J. Struct. Biol. **135**, 239–250 (2001)
7. Kopp, J.: Efficient numerical diagonalization of hermitian $3 \times 3$ matrices. Int. J. Mod. Phys. C **19**, 523–548 (2008)
8. Lucic, V., Rigort, A., Baumeister, W.: Cryo-electron tomography: the challenge of doing structural biology in situ. J. Cell Biol. **202**, 407–419 (2013)
9. Moreno, J.J., Martínez-Sánchez, A., Martínez, J.A., Garzón, E.M., Fernández, J.J.: TomoEED: fast edge-enhancing denoising of tomographic volumes. Bioinformatics **34**(21), 3776–3778 (05 2018). https://doi.org/10.1093/bioinformatics/bty435
10. Peddie, C.J., Collinson, L.M.: Exploring the third dimension: Volume electron microscopy comes of age. Micron **61**, 9–19 (2014)
11. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C. The art of scientific computing, 2nd ed. Cambridge University Press, Cambridge (2002)

[1] http://www.cnb.csic.es/ %7ejjfernandez/tomoeed

# El Modelo Dispositivo - Interacción y Machine Learning en Interacción Natural

Juan Jesús Ojeda Castelo[1]

Applied Computing Group (TIC-211), Departamento de Informática, Universidad de Almería[1]
juanje.ojeda@ual.es

**Abstract.** En esta Tesis se está realizando una investigación sobre cómo obtener un modelo basado principalmente en interacción natural. Dicha interacción tiene la peculiaridad de que es difícil de conseguir una verdadera interacción intuitiva para el usuario debido a que necesita dos características fundamentales: El tiempo de respuesta sea el mínimo posible y la fiabilidad sea muy alta. En general, los métodos utilizados para el reconocimiento de gestos se definen porque si tienen una fiabilidad alta el tiempo de respuesta también es mayor y por el contrario si el tiempo de respuesta es óptimo, su rendimiento se ve afectado negativamente. En este informe se explica cómo se está progresando en este aspecto. Por otro lado, se describe como se está abordando otro objetivo importante de esta Tesis, como es el aspecto de la adaptación cuando se integra algún modo de interacción natural en el sistema. En este caso se ha utilizado modelos de usuario y la creación del modelo dispositivo-interacción para mejorar la adaptación en relación a la interacción con un sistema informático. Los resultados de este experimento son prometedores.

## 1 Introducción

En el enfoque tradicional del campo de la Visión Artificial se buscan características que permitan identificar los elementos que se quiere detectar previamente establecidos [1]. Ejemplos de estas características son: esquema de colores, textura de la imagen o simplemente esquinas.

La relación entre Machine Learning y Visión Artificial puede ser descrita como que Machine Learning ayuda a entender lo que la Visión Artifical ve. Esta forma de describir de una manera sencilla el rol que interpretan ambas disciplinas, quiere decir que la Visión Artificial se centra en técnicas que permitan detectar elementos en una imagen o vídeo [2] y realizar segmentación en este tipo de formatos multimedia [3]. Sin embargo, la competencia de la Visión Artificial se queda ahí, pero Machine Learning aporta a este proceso la habilidad para aprender e ir mejorando el reconocimiento o realizar una clasificación de los elementos detectados y de esta forma saber exactamente qué objeto es, entre otras funciones.

Deep learning es una disciplina que ha tenido un gran impacto en el ámbito de Visión Artificial gracias a las Redes Neuronales Convolucionales [4, 5]. Un

ejemplo de esta afirmación es el desafío ILSVRC-2012, el cual se basa en un dataset denominado ImageNet con 1,2 millones de imágenes de entrenamiento a alta resolución. En este reto se tiene que etiquetar cada una de las imágenes con el objeto que representa a las mismas. El ganador de la competición fue una Red Convolucional denominada AlexNet.

Por otro lado, la adaptación es un factor relevante/útil en la interacción del usuario puesto que cada individuo posee unas características particulares. La adaptación en términos de interacción puede hacer que una persona que tenga movilidad reducida sea capaz de utilizar un sistema informático o una persona que solo pueda mover el brazo izquierdo interaccione exclusivamente con dicho brazo. En general, la arquitectura de los sistemas cuyo objetivo es adaptarse al usuario, está compuesta por tres elementos: Modelos de usuario, modelos de dominio y modelos de adaptación [6]. El modelo de adaptación describe cómo debe de estar dispuesto el contenido para el usuario en la capa de presentación a partir de la información contenida en los modelos de usuario y de dominio.

El modelo de usuario es un componente que sirve para adaptar y personalizar un sistema al usuario debido a la cantidad de información que contiene respecto al mismo [7]. Estos modelos representan los objetivos, el conocimiento, los interes y otras características que se consideren significativos del usuario. Estos modelos se han estudiado y se han aplicado para representar a la persona que interacciona con el sistema y entre otros objetivos, poder mejorar la usabilidad del mismo. Algunos de los ámbitos en los que se está aplicando estos modelos son: Redes sociales [8], motores de búsqueda [9] y sistemas de recomendación [10].

En la sección 2 se describe el avance de la investigación en términos de adaptación y reconocimiento de gestos con la mano para conseguir una interacción natural. Por último, en la sección 3 se expone la resolución que ha obtenido el autor en su investigación durante el último año.

## 2 Progreso de la Investigación

En el anterior informe se propuso la idea de implementar un algoritmo basado en características para realizar la detección de las manos pero no se han obtenido los resultados esperados con este algoritmo así que se realizó una pausa en este aspecto para investigar más en el tema y escoger otra solución que pueda ofrecer un mejor rendimiento que será explicada más adelante.

Mientras se decidía otra alternativa para el proceso de reconocimiento de gestos se estuvo progresando en el objetivo de la adaptación que concierne a esta Tesis. Para avanzar en dicho objetivo se han combinado los modelos de usuario con un modelo dispositivo-interacción.

La arquitectura del sistema propuesto se compone de cuatro elementos fundamentales: El módulo de interacción, modelo de usuario, modelo dispositivo-interacción y módulo de adaptación.

El **módulo de interacción** se encarga de obtener los datos del dispositivo, en este caso de Microsoft Kinect v2, ya que dicho dispositivo es capaz de transmiter diversos flujos de datos: profundidad, audio, cámara de color. Este

módulo se encarga de procesarlos y que puedan ser utilizados por es sistema donde dependiendo del tipo de interacción que se vaya a realizar servirán para el reconocimiento de gestos o simplemente para detectar el movimiento de partes del cuerpo del individuo.

El **modelo de usuario** que se ha integrado en el sistema es del tipo basado en características, el cual está compuesto por una serie de propiedades del usuario. En este caso se han elegido los siguientes atributos para representar a un usuario:

- Nombre y Apellidos.
- Edad.
- Sexo.
- Problemas de lateralidad.
- Discapacidad.

El **modelo dispositivo-interacción** es la principal aportación al módulo de adaptación, siendo su objetivo principal proporcionar a la interacción los términos de adaptación que se requieren en esta fase. Este modelo se centra en las características del dispositivo para optimizar al máximo la interacción, en función del usuario. Las características de este modelo son:

- **Detectar si está de pie o sentado**: Esta opción está pensada para facilitar la interacción a usuarios que estén en silla de ruedas y no puedan ponerse en pie. De esta forma se tiene en cuenta solo la parte superior del cuerpo y se ignora la parte inferior.
- **Activar o desactivar la cámara RGB**: Está diseñado especialmente para personas que tienen un nivel cognitivo bajo y necesitan tener activada la cámara para verse en espejo y coordinar los movimientos.
- **La distancia según el sensor de profundidad**: Almacena la distancia a la que tiene que coocarse el usuario respecto al dispositivo para que tenga una óptima interacción con el entorno.
- **Movilidad de los brazos**: Es necesario comprobar si el usuario tiene movilidad completa o por el contrario no puede mover alguno de sus brazos para que la interacción solo se centre en la extremidad que tiene movilidad e ignore el resto.

El último elemento de esta arquitectura está formado por un conjunto de reglas de adaptación que son evaluadas a partir de la información contenida tanto en el modelo de usuario como en el modelo dispositvo-interacción para ofrecer la acción correspondiente.

Con el fin de demostrar la validez de este modelo se ha realizado un experimento, en el cual se han visto involucrados doce participantes cada uno con una afección diferente: discapacidad visual, discapacidad auditiva, discapacidad física y autismo. Estos usuarios realizaron una serie de tareas durante dos iteraciones, cada una compuesta de tres sesiones donde se contabilizaron el tiempo de realización y el número de errores cometidos. Ambos parámetros fueron disminuyendo conforme los participantes realizaban más sesiones.

Después de la realización de este experimento le dediqué más tiempo a investigar la forma de realizar un reconocimiento de gestos que fuera lo más natural posible. Durante este período leí multitud de fuentes de información y sobre todo actuales, sobre qué se está haciendo en el campo del reconocimiento de gestos. La corriente deriva a utilizar los campos de conocimiento de la Inteligencia Artificial que involucran aprendizaje, es decir, técnicas que permiten aprender al sistema: Machine Learning y Deep Learning. Con esta premisa decidí utilizar algoritmos de estas especialidades para realizar el reconocimiento de gestos. En la actualidad hay cinco técnicas que son muy relevantes en este ámbito:

- Redes Bayesianas.
- Redes Neuronales.
- Algoritmos genéticos.
- Lógica difusa.
- Redes Neuronales Convolucionales (Deep Learning).

En este momento se está desarrollando la interacción natural con el uso de las técnicas mencionadas anteriormente aunque no hay resultados concluyentes para afirmar cuál de ellas es la mejor opción para el problema planteado. Sin embargo, de acuerdo a las pruebas que se han realizado hasta la fecha puedo comentar que quizás el desenlace de este paradigma sea la implementación de una técnica híbrida. Es necesario añadir que el principal inconveniente que se puede encontrar en estas técnicas es la cantidad elevada de datos que se necesitan a priori para entrenar el sistema, especialmente en Deep Learning que usualmente los datos necesarios entran en la categoría de Big Data.

## 3 Conclusiones

La aplicación de algoritmos que generalmente han sido utilizados en el enfoque tradicional de Visión Artificial para la detección de objetos, no han mostrado la eficiencia esperada para el reconocimiento de las manos. La problemática principal es que si el fondo de la imagen tiene muchos elementos (imaginemos una habitación con todos los muebles y objetos que suele haber), además de las características o "keypoints" que obtiene de las manos también va a detectar las caracterísitcas de los objetos que haya en el entorno y esta situación hace extremedamente difícil discernir entre las características de las manos y las del resto de elementos.

Este contexto ha tenido como consecuencia que se busque otra alternativa que se adapte mejor a los objetivos de esta Tesis. Por esta razón se ha decidido incluir Inteligencia Artificial, y más concretamente Deep Learning y Machine Learning porque están dando muy buenos resultados en el campo de Visión Artificial [11]. En este momento se están probando diferentes técnicas para comprobar su validez en el modelo propuesto pero no se han obtenido suficiente resultados para determinar una conclusión.

En el tema de la adaptación se ha diseñado un modelo dispositivo-interacción el cual se centra en las características del dispositivo para ofrecer una interacción

más adaptable al usuario. La combinación de dicho modelo y el modelo de usuario propuesto resulta útil para la adaptación de la interacción del usuario. Esta afirmación está basada en el experimento realizado con usuarios que poseían diferentes discapacidades (visual, auditiva, física y autismo), el cual muestra que la tasa de errores de las tareas propuestas han disminuido con el tiempo y que omitiendo el modelo de dispositivo-interacción o cambiando las características óptimas establecidas para el individuo en el sistema, muchos de ellos no eran capaces de completar las tareas propuestas.

## References

[1] Lee, A.: Comparing deep neural networks and traditional vision algorithms in mobile robotics. Swarthmore University (2015)

[2] Shin, M.C., Goldgof, D.B., Bowyer, K.W.: Comparison of edge detector performance through use in an object recognition task. Computer Vision and Image Understanding **84**(1) (2001) 160–178

[3] Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1369–1378

[4] Ramírez, I., Cuesta-Infante, A., Pantrigo, J.J., Montemayor, A.S., Moreno, J.L., Alonso, V., Anguita, G., Palombarani, L.: Convolutional neural networks for computer vision-based detection and recognition of dumpsters. Neural Computing and Applications (2018) 1–9

[5] Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Hasan, M., Van Esesn, B.C., Awwal, A.A.S., Asari, V.K.: The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164 (2018)

[6] Knutov, E., De Bra, P., Pechenizkiy, M.: Ah 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. New Review of Hypermedia and Multimedia **15**(1) (2009) 5–38

[7] Brusilovsky, P.: From adaptive hypermedia to the adaptive web. In: Mensch & Computer 2003. Springer (2003) 21–24

[8] Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In: International Conference on User Modeling, Adaptation, and Personalization, Springer (2011) 1–12

[9] Cramer, M., Zhai, C., Shen, X., Tan, B.: Real time implicit user modeling for personalized search (May 14 2013) US Patent 8,442,973.

[10] Jawaheer, G., Weller, P., Kostkova, P.: Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. ACM Transactions on Interactive Intelligent Systems (TiiS) **4**(2) (2014) 8

[11] Shanmugamani, R., Sadique, M., Ramamoorthy, B.: Detection and classification of surface defects of gun barrels using computer vision and machine learning. Measurement **60** (2015) 222–230

# Optimizando la eficiencia energética de SMACOF

Francisco José Orts Gómez

Grupo de Supercomputación-Algoritmos, Dpt. de Informática, Univ. de Almería, ceiA3, 04120, Almería, España.
francisco.orts@ual.es

**Resumen** Reducir la dimensionalidad de un conjunto grande de datos es de especial interés en muchos campos, tales como psicología, sociología o análisis de mercados. De entre todas las técnicas existentes para la reducción de dicha dimensionalidad, son especialmente utilizados los denominados métodos de escalamiento multidimensional (MDS de sus siglas en inglés). Estos métodos permiten mapear los datos originales desde un espacio de muchas dimensiones a otro espacio con menos. Sin embargo, los métodos MDS consumen una gran cantidad de recursos computacionales, por lo que actualmente existe un gran interés en optimizarlos, habiendo una amplia investigación al respecto. SMACOF es uno de los métodos de MDS más precisos y que también tiene un mayor coste computacional. En este trabajo se han implementado y evaluado dos versiones eficientes de SMACOF, una versión multicore y otra basada en GPUs.

**Keywords:** SMACOF · escalamiento multidimensional · reducción de dimensionalidad · eficiencia energética.

## 1. Introducción

Los métodos de reducción de dimensionalidad tienen por objetivo mapear datos pertenecientes a un espacio de grandes dimensiones a uno con menos dimensiones, algo que supone una optimización especialmente importante de cara a poder manejar esos datos más eficientemente. Una de las aplicaciones que tiene esta reducción de dimensionalidad es el poder ver gráficamente la estructura de los datos de alta dimensionalidad en un espacio 2D o 3D para facilitar la comprensión de los mismos. De entre los diversos métodos ´existentes para reducir la dimensionalidad, los denominados métodos de Escalamiento Multidimensional (MDS) son muy populares [5]. En esta línea se pueden encontrar varias aplicaciones en [9], [14], [16]. Además, MDS ha demostrado ser especialmente útil como técnica para evaluar los criterios de clasificación de objetos [11] o descubrir tendencias de los datos que inicialmente habían pasado desapercibidas [4], siendo útil como un modelo psicológico que permite descubrir patrones humanos de conduca [12].

SMACOF (Scaling by MAjorizing a COmplicated Function) es uno de los algoritmos más conocidos de MDS [7], siendo el más preciso [13] pero también

el más costoso, ya que su complejidad es $O(m^2)$, siendo $m$ el número de observaciones. En un esfuerzo por reducir dicho coste, [17] redujo su complejidad a $O(m\sqrt{m})$ desarrollando un modelo iterativo. En [23], los autores redujeron la complejidad a $O(m \log m)$ particionando la matriz original en submatrices y luego combinando las soluciones de cada submatriz para obtener una solución global. No obstante, la complejidad de estas versiones más optimizadas y en general de cualquier técnica de MDS sigue siendo grande [13], por lo que deben considerarse estrategias de paralelización para acelerar su cálculo [18,19].

Actualmente, entre los objetivos de la HPC se incluye la optimización del consumo de energía. La relación entre la velocidad computacional y la potencia eléctrica (GFLOPs/ watt) se puede utilizar como un indicador adecuado de eficiencia energética [15]. Aumentar este parámetro significa que el sistema logra un mejor rendimiento (GFLOP) con menos energía eléctrica (vatios), dando como resultado un consumo menor de energía. Por lo tanto, el ratio debe ser maximizado. En este trabajo, dos versiones paralelas del algoritmo SMACOF, una para multinúcleo y otra para GPU, se presentan y evalúan en diferentes arquitecturas modernas, teniendo en cuenta no solo la velocidad de computo, sino también la eficiencia energética.

## 2. El algoritmo SMACOF para MDS

El escalamiento multidimensional es una técnica para el análisis de similitudes o disimilitudes de datos (items). Su objetivo es encontrar una serie de puntos $Y_1, Y_2, \ldots, Y_m \equiv Y$ en el espacio de dimensiones reducidas $\mathbb{R}^s$, $s < n$, de forma que las distancias entre ellos sean lo más parecidas posibles a las distancias entre los puntos originales $X_1, X_2, \ldots, X_m \equiv X$ en el espacio $\mathbb{R}^n$.

La atención de este trabajo se centra en el algoritmo SMACOF, cuyo objetivo es minimizar la función de estrés [7]. SMACOF ha demostrado mejores resultados al optimizar la función de estrés en comparación con otras propuestas [13]. La idea principal se basa en el concepto de mayorización, que consiste en aproximar una función compleja mediante otra más simple. Este método busca iterativamente una nueva función, que se encuentra por encima la función original, tocándola en un punto de soporte. En cada iteración, el mínimo de la nueva función está más cercano al mínimo de la función original, que en este caso es la función de estrés [5]. SMACOF puede ser expresado como se muestra en el Alg. 1.

El algoritmo 1 tiene altos costes computacionales y de consumo de memoria debido a las grandes estructuras de datos involucradas: la matriz de entrada $\Delta$ $(m \times m)$, la de salida y las matrices auxiliares $(m \times s)$ y otras tres matrices auxiliares $(m \times s)$ para almacenar las disimilitudes entre los items del espacio de dimensiones reducidas. La simetría no se ha explotado a la hora de almacenar las estructuras de datos, pero sí se ha tenido en cuenta para la actualización de las matrices. Teniendo esto cuenta, el número de operaciones del Alg. 1 es: $3s/2m^2 + 3s/2m$ para la inicialización (línea 2 de Alg. 1) y $(7/2s + 3/2)m^2 + 1/2(3s + 1)m$ para el proceso iterativo.

**Algoritmo 1** SMACOF($m$, $s$, $\Delta$, $kmax$, $\epsilon$, $Y$)

**Require:**
$m$: número de items;
$s$: número de dimensiones del espacio reducido;
$\Delta$: $m \times m$ matriz de disimilitudes en el espacio original;
$kmax$: número máximo de iteraciones;
$\epsilon$: límite para la varianza del estrés
**Ensure:**
$Y$: conjunto de puntos de búsqueda en el espacio reducido en una matriz $m \times s$
1: Generar solución inicial aleatoriamente, $Y^0$
2: **Calcular distancias euclideas**, $D^0 = [d(Y_i^0, Y_j^0)]$ ▷ $O(m^2 s)$
3: $k = 0$, $error = 1$
4: **if** $(k < kmax)$ and $(error > \epsilon)$ **then**
5: **Calcular la matriz de la transformada de Guttman**, $B^k \equiv B^k(\Delta, D^{k-1})$ ▷ $O(m^2)$
6: **Calcular la transformada de Guttman**, $Y^k = 1/m \cdot B^k \cdot Y^{k-1}$ ▷ $O(m^2 s)$
7: **Actualizar distancias** $D^k = [d(Y_i^k, Y_j^k)]$ ▷ $O(m^2 s)$
8: **Calcular** $E_{MDS}^k$
9: $error = |E_{MDS}^k - E_{MDS}^{k-1}|$
10: $k = k + 1$
11: **return** $Y$

## 3. Implementaciones paralelas del agoritmo SMACOF

El coste computacional de SMACOF es $O(s \cdot m^2)$ y los requisitos de memoria son del orden $O(m^2)$. Estos valores han limitado durante años la aplicabilidad de SMACOF para resolver grandes problemas de MDS. Por suerte, el uso de técnicas HPC puede ayudar a superar estos inconvenientes. En este trabajo, proponemos dos versiones paralelas basadas en la explotación de arquitecturas multinúcleo y GPUs modernas respectivamente. Esta sección está dedicada a describir estas implementaciones.

Ambas implementaciones se centran en la ejecución paralela del cálculo de las matrices de distancias euclideas (líneas 2 y 7 de Alg. 1) y de la transformada de Guttman (líneas 5 y 6 de Alg. 1). Los procedimientos paralelos se resaltan en negrita en el Alg. 1. Para calcular los resultados de estos procedimientos, hemos tenido en cuenta que se trabaja con matrices simétricas ($B^k$, $D^k$ and $\Delta$). Por ejemplo, para calcular la matriz simétrica $B^k$ (que define la transformada de Guttman) solo es necesario calcular una submatriz triangular de $L = (m \cdot (m + 1)/2)$ elementos. Por lo tanto, $B^k$ se puede manipular como un vector unidimensional de $L$ elementos que se pueden actualizar en paralelo. De esta forma, dos bucles anidados se integran en un solo bucleque que calcula la matriz triangular de $L$ elementos, lo que permite paralelizar fácilmente de forma que se mantenga el balanceo de carga. Esta idea también ha sido aplicada al cómputo de $D^k$.

La versión multinúcleo se ha implementado utilizando C, OpenMP [6] y la librería MKL [3]. El cálculo de $B^k$ y $D^k$ tiene en cuenta la simetria de estas

matrices. El primer bucle de dicho algoritmo se distribuye entre los núcleos y cuando termina se incluye un punto de sincronización para garantizar que los elementos no diagonales de $B^k$ se calculan antes de comenzar el segundo bucle. Por otro lado, la librería MKL (concretamente la rutina cblas_dgemm) se encarga de calcular en paralelo el producto matriz-matriz vinculado a la transformada de Guttman (línea 6 de la Alg. 1).

En la versión GPU, se han codificado tres kernels utilizando C y CUDA para calcular en paralelo $D^k$ (líneas 2 y 7 de Alg. 1) y $B^k$ (línea 5 de Alg. 1). Las distancias euclideas requieren un kernel, y la transformada de Guttman requiere dos, como se explica a continuación. Para calcular la matriz de distancias, cada hilo actualiza dos elementos simétricos de la matriz $D^k$. Por otra parte, se han utilizado intruccions shuffle para las reducciones implicadas en el cálculo de elementos de $D^k$. Estas instrucciones, disponibles a partir de la arquitectura Kepler de NVIDIA, permiten que los hilos del mismo warp compartan información [2]. En las pruebas realizadas para este trabajo, las instrucciones shuffle han demostrado mejoras en el rendimiento en comparación con las reducciones basadas en memoria compartida. Hemos observado que la ventaja de las instrucciones shuffle frente a la versión de memoria compartida aumenta con $s$. Específicamente, se ha evaluado el rendimiento para tamaños de problema de $m = 10000$ a $m = 40000$ con $s = 64$ y la versión shuffle ha obtenido el mismo o mejor rendimiento (hasta 30 %) que la versión de memoria compartida en el cálculo de la matriz $D^k$ (líneas 2 y 7 de Alg. 1).

La versión CUDA usa dos kernels para calcular $B^k$. En el primer kernel, cada hilo comienza calculando un elemento no diagonal de $B^k$. A continuación, su simétrico se copia sin necesitarse ninguna sincronización. Cuando este primer kernel finaliza, el segundo calcula los elementos diagonales a partir de los que no son diagonales. Para ello se utiliza la rutina *cublasDgemm* de cuBLAS [1] para acelerar el producto matriz-matriz en GPU (línea 6 de Alg. 1).

## 4. Ajuste de la eficiencia energética del algoritmo SMACOF

La Eficiencia Energética (EE) generalmente se define como la relación entre la velocidad computacional y la potencia eléctrica, es decir, *GFLOPs/watt* [15]. Por lo tanto, para una ejecución óptima, la relación debe maximizarse.

La optimización de la EE puede verse como un problema de planificación de máquinas paralelas con coste [22]. Las versiones paralelas de SMACOF se pueden ejecutar en una de las plataformas disponibles. Cada plataforma se denota como $\mathcal{F}^k \in \mathcal{F}$, $k = 1, \ldots, f$ donde $\mathcal{F}$ es el conjunto de las $f$ plataformas disponibles. Cada plataforma $\mathcal{F}^k$ consiste en un conjunto de máquinas paralelas $\mathcal{M}^k$, $\mathcal{F}^k = \{\mathcal{M}_i^k\}_{i=1}^{c_k}$, siendo $c_k$ el número de máquinas que tiene $k$. La eficiencia energética correspondiente depende del número de máquinas involucradas en el cálculo y el tamaño de los datos.

Entonces, la solución al problema de planificación es un subconjunto de plataformas $\mathcal{F}^{k_o} \subseteq \mathcal{F}$ con sus configuraciones óptimas definidas por las máquinas

$r^o_{k_o}$ que optimizan la EE ($r^o_{k_o} \leq c_{k_o}$). Proponemos un enfoque heurístico para resolver este problema, basado en un modelo funcional de EE para plataformas modernas (multinúcleo y GPU).

Los modelos para estimar la EE deben combinar rendimiento y potencia. Algunos trabajos previos han propuesto modelos funcionales para la estimación de la EE en aplicaciones iterativas [10]. Centrándose en una ejecución particular de la aplicación con operaciones de punto flotante en una plataforma homogénea $k$, y suponiendo un balanceo de carga perfecto entre las máquinas activas $r_k$, entonces el siguiente modelo de EE en función de $r_k$ es razonable:

$$EE(r_k) = \frac{F}{(\frac{T^k(1)}{r_k} + TC^k(r_k))(P^k_{idle} + r_k p^k(r_k))} \quad (1)$$

donde $T^k(r_k)$ y $P^k(r_k)$ son el tiempo de ejecución y el consumo de energía en $r_k$ maquinas respectivamente, $TC^k(r_k)$ representa la penalización en el tiempo de ejecución debido a la contención entre las máquinas activas en la plataforma $k$, $P^k_{idle}$ representa el consumo de energía inactivo cuando ningún proceso está utilizando activamente una máquina y $p^k(r_k)$ es la contribución de potencia de cada máquina

Según este modelo $T^k(r_k)$, se obtiene un mínimo para un número de máquinas activas ya que $TC^k(r_k)$ es una función creciente y $P^k(r_k)$ también es una función creciente para $r_k$. Por lo tanto, $EE(r_k)$ alcanza un máximo para las máquinas $r^o_{k_o}$.

Por lo tanto, desde el punto de vista del uso de SMACOF, para optimizar la EE se debe identificar $r^o_k$ en el conjunto de plataformas disponibles y elegir la plataforma $k_o$ que optimiza la EE, es decir, logra que EE($r^o_{k_o}$). Las computadoras modernas ofrecen esencialmente dos tipos de hardware, procesadores multinúcleo y GPUs, siendo lo normal que una plataforma esté constituida por los dos tipos. Hemos definido una heurística para decidir cuál es la mejor plataforma para ejecutar las los versiones de SMACOF. Nuestra propuesta se organiza en dos etapas: primero, identificar la configuración óptima de cada plataforma, y segundo, seleccionar la plataforma y configuración óptimas. Las consideraciones previas sobre el modelo de EE ayudan a definir una exploración eficiente de evaluación del rendimiento para encontrar las configuraciones óptimas en cada plataforma. Por lo tanto, la búsqueda selectiva descrita en el Algoritmo 2 se puede utilizar para encontrar las plataformas óptimas y sus configuraciones.

## 5. Evaluación

En esta sección, se evalúa el algoritmo SMACOF en términos de tiempo de ejecución y eficiencia energética en tres plataformas diferentes:

$\mathcal{F}_1$ : Bullion S8: 4 Intel Xeon E7 8860v3 (16 × 4 núcleos);
$\mathcal{F}_2$ : Bullx R421-E4 Intel Xeon E5 2620v2 (12 núcleos y 64 GB RAM);
$\mathcal{F}_3$ : NVIDIA K80 (compuesta por dos GPUs Kepler GK210) conectadas a un host Bullx R421-E4 Intel Xeon E5 2620v2.

**Algoritmo 2** Heurística para calcular el conjunto de plataformas óptimas $\{k_o\}$, con sus configuraciones $\{r^o_{k_o}\}$, que optimizan la EE de SMACOF

**Require:**
$\mathcal{F} = \{\mathcal{F}^k\}_{k=1}^{f}$ with $\mathcal{F}^k = \{\mathcal{M}_i^k\}_{i=1}^{c_k}$; ▷ Conjunto de plataformas
Versiones paralelas de SMACOF($m$, $s$, $\Delta$, $kmax$, $\epsilon$, $Y$) para ejecutar en las $f$ plataformas disponibles;
$m$ (items), $s$ (dimensiones de salida); ▷ Tamaño de los datos
*sampling*.
**Ensure:**
$\{k_o, r^o_{k_o}\}$ optimiza la EE en las $f$ plataformas disponibles
1: Evalua el número de operaciones de punto flotante de SMACOF($m$, $s$, $\Delta$, $kmax$, $\epsilon$, $Y$)
2: **for** $k \leftarrow 1$ **to** $f$ **do**
3: Ejecuta SMACOF($m$, $s$, $\Delta$, $kmax$, $\epsilon$, $Y$) en $r_k = c_k$ máquinas y evalúa su EE denotado como $\mathcal{EE}^k$
4: **for** $i \leftarrow c_k - sampling$ **to** $sampling$ **do**
5: Ejecuta SMACOF($m$, $s$, $\Delta$, $kmax$, $\epsilon$, $Y$) en $r_k = i$ máquinas y evalúa $\mathcal{EE}^{Aux}$
6: **if** $\mathcal{EE}^{Aux} \leq \mathcal{EE}^k$ **then**
7: $r^o_k = i + sampling$
8: Break blucle-$i$
9: **else**
10: $\mathcal{EE}^k = \mathcal{EE}^{Aux}$
11: Selecciona las plataformas $\{k_o\}$ con sus configuraciones óptimas $\{r^o_{k_o}\}$ que maximizan la EE
12: **return** $\{k_o, r^o_{k_o}\}$

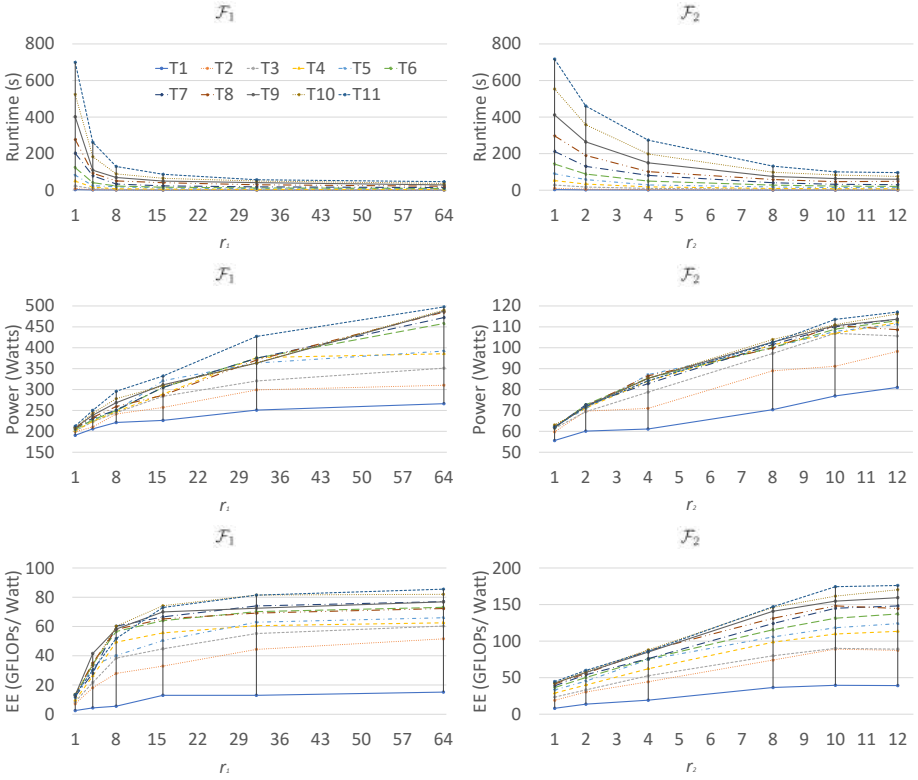
$\mathcal{F}_1$, $\mathcal{F}_2$ y $\mathcal{F}_3$ ejecutan Ubuntu 16.04 LTS y $\mathcal{F}_3$ ejecuta CUDA Toolkit 8. Los programas han sido compilados usando gcc 5.4.0 y nvcc 8.0.44 con flags de optimización O3 para arquitectura de GPU 3.5. Para la adquisición de datos de medición de energía, hemos recopilado esta información de varios contadores de hardware. Para Intel, hemos utilizado la interfaz Running Average Power Limit (RAPL) y, para NVIDIA, la librería de administración de NVIDIA (NVML).

Para la evaluación de SMACOF se han considerado problemas de diferentes tamaños definidos por los valores de $m$, $n$ y $s$ (ver Tabla 1). Se han usado datos de entrada generados aleatoriamente. Asimismo, el número de iteraciones evaluadas ha sido 100.

**Tabla 1.** Batería de pruebas realizada variando el número de items ($m$), de dimensiones del espacio original ($n$), y dimensiones del espacio reducido ($s$).

| | $T1$ | $T2$ | $T3$ | $T4$ | $T5$ | $T6$ | $T7$ | $T8$ | $T9$ | $T10$ | $T11$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $m$ | 2000 | 4000 | 6000 | 8000 | 10000 | 12000 | 14000 | 16000 | 18000 | 20000 | 22000 |
| $n$ | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1100 |
| $s$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |

**Figura 1.** Tiempo de ejecución (arriba), potencia (medio) y eficiencia energética (abajo) de las pruebas realizadas en la Table 1 en las plataformas $\mathcal{F}_1$ y $\mathcal{F}_2$. Se ha utilizado un color diferente para cada una de las columnas de la tabla mencionada, es decir, para cada configuración diferente.




La Figura 1 muestra el tiempo de ejecución, la potencia y la eficiencia energética del conjunto de prueba en las plataformas $\mathcal{F}_1$ y $\mathcal{F}_2$ (multinúcleo) y la tabla 2 muestra los mismos parametros para la plataforma $\mathcal{F}_3$ (GPU) para esas mismas prueba. El tiempo de ejecución de las versiones multinúcleo (representadas en la parte superior de la Fig. 1) cumple con lo dicho para los modelos descritos en la Sección 4. El tiempo de ejecución disminuye con los valores de $r_1$ y $r_2$, por lo tanto se obtiene el mejor rendimiento para la cantidad máxima de núcleos. Las mediciones de potencia se muestran en la parte media de la Fig. 1. Se hace evidente al observar la figura que la evolución temporal de la potencia depende parcialmente de factores impredecibles para los programadores. Para superar este problema, ha sido necesario recopilar las mediciones tras un período de actividad del procesador para minimizar su varianza debido a cambios de la temperatura. Esta inestabilidad se puede observar en el gráfico de potencia para ambas plataformas, pero podemos concluir que la tendencia

**Tabla 2.** Tiempo de ejecución, potencia y eficiencia energética de las pruebas realizadas en la Tabla 1 en las plataformas $\mathcal{F}_3$ (GPU).

| $\mathcal{F}_3$ | $T1$ | $T2$ | $T3$ | $T4$ | $T5$ | $T6$ | $T7$ | $T8$ | $T9$ | $T10$ | $T11$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Time (s) | 2.8 | 4.9 | 5.1 | 5.9 | 6.5 | 11.0 | 18.8 | 28.7 | 42.3 | 64.9 | 91.5 |
| Power (Watts) | 38.7 | 98.6 | 105.2 | 108.0 | 112.6 | 113.6 | 112.8 | 110.1 | 112.3 | 111.5 | 110.3 |
| EE (GFLOPs/ Watt) | 13.8 | 24.7 | 75.1 | 150.0 | 255.1 | 257.2 | 240.7 | 241.7 | 228.7 | 205.9 | 196.6 |

**Tabla 3.** Valores de EE de la prueba $T11$ obtenidos mediante la heurística del Alg. 2 para las plataformas multinúcleo $\mathcal{F}_1$ y $\mathcal{F}_2$.

| $\mathcal{F}_1$ | | |
|---|---|---|
| $r_1$ | 64 | 61 |
| EE (GFLOPs/Watt) | 85.5 | 85.0 |

| $\mathcal{F}_2$ | | |
|---|---|---|
| $r_2$ | 12 | 9 |
| EE (GFLOPs/Watt) | 176.1 | 155.8 |

del consumo de energía aumenta a medida que lo hace el número de núcleos y el tamaño del problema.

Centrando la atención en la eficiencia energética (mostrada en la parte inferior de la Fig. 1), se ve que esta aumenta a medida que lo hace el número de núcleos. Los valores más altos de $r_1$ y $r_2$ optimizan la eficiencia energética. Por lo tanto, el valor óptimo de $r_k$ en ambas plataformas se encuentra en un intervalo amplio, por ejemplo 32-64 (10-12) para $\mathcal{F}_1$ ($\mathcal{F}_2$).

Para elegir la plataforma óptima, se pueden comparar las tres plataformas en términos de rendimiento. De esta forma, la mejor opción para $T11$ es $\mathcal{F}_1$ ya que los tiempos de ejecución son 46.6s, 96.2s y 91.5s en $\mathcal{F}_1$ con $r_1^o = 64$, $\mathcal{F}_2$ con $r_2^o = 12$ y $\mathcal{F}_3$ respectivamente. Esta selección es la misma para todos los casos de prueba. Si nos enfocamos en la eficiencia energética, la mejor opción es la GPU cuando el tamaño del problema es suficientemente grande ya que se consume menos energía que en $\mathcal{F}_1$ y logra un rendimiento razonable. Así pues, para optimizar la eficiencia energética, la mejor opción es el uso de la plataforma GPU para los casos de prueba $T4 - T11$. Por ejemplo, para $T11$, la eficiencia energética en las diferentes plataformas es 85.5, 176.1 y 196.6 GFLOP/ watt para $\mathcal{F}_1$, $\mathcal{F}_2$ y $\mathcal{F}_3$, respectivamente. La mejor plataforma para $T1 - T3$ es la multinucleo $\mathcal{F}_2$ ya que consume menos energía que $\mathcal{F}_1$.

Estos resultados respaldan el proceso de evaluación explicado en la Sección 4 para explorar de forma automática la selección de la plataforma óptima y de recursos. Este procedimiento ha sido desarrollado en Python. Hemos elegido $sampling = 3$ para obtener diferencias relevantes entre las evaluaciones de las plataformas. Los resultados respaldan la idea de comenzar el proceso de evaluación por el mayor número de núcleos de CPU disponibles en cada plataforma e ir reduciendo hasta encontrar el $r_k$ óptimo. Para ilustrar la evaluación del rendimiento (Alg. 2) para las plataformas multinúcleo, nos centramos en la prueba $T11$. La tabla 3 muestra la EE obtenida cuando se ejecuta un conjunto de diez iteraciones de SMACOF en las plataformas $\mathcal{F}_1$ y $\mathcal{F}_2$. Solo se requieren dos muestras para la exploración de la evaluación del rendimiento para $T11$, ya que

$r_1^o = 64$ y $r_2^o = 12$ se identifican por el preproceso. Podemos concluir que el modelo para evaluar y escoger la plataforma óptima funciona adecuadamente.

## 6. Conclusiones

Este trabajo ha analizado un enfoque para optimizar la eficiencia energética (GFLOPs/watt) del algoritmo SMACOF, un método conocido y preciso para resolver problemas con MDS. Se han desarrollado y evaluado dos versiones paralelas de SMACOF, multinúcleo y GPU, así como un software en Python complementario basado en en un enfoque heurístico para explorar la configuración óptima para las dos versiones de SMACOF.

Se ha llevado a cabo una evaluación experimental en tres plataformas: con 64 núcleos, 12 núcleos y con una GPU. Los resultados muestran que la plataforma de 64 núcleos es la mejor plataforma para optimizar el tiempo de ejecución de SMACOF; la plataforma de 12 núcleos es la mejor opción para mejorar la eficiencia energética para problemas reducidos y, para los problemas más grandes, la eficiencia energética óptima se alcanza con la GPU.

En las versiones de SMACOF disponibles actualmente sólo se ha considerado el tiempo de ejecución, y no se optimizan ni el consumo de energía ni la capacidad de adaptación a la plataforma y al tamaño del problema. Por lo tanto, nuestras implementaciones de SMACOF son de gran interés para el desarrollo de aplicaciones que tengan en cuenta la eficiencia energética. Como trabajo futuro, estamos considerando implementar una versión paralela distribuida de SMACOF y analizar otros métodos para resolver problemas de MDS.

## Referencias

1. cuBLAS library (2017). URL http://docs.nvidia.com/cuda/cublas/index.html
2. CUDA Pro Tip: Do The Kepler Shuffle (2017). URL https://devblogs.nvidia.com/parallelforall/cuda-pro-tip-kepler-shuffle/
3. Intel Math Kernel Library (Documentation) (2017). URL https://software.intel.com/en-us/mkl/documentation
4. Bilsky, W., Borg, I., Wetzels, P.: Assessing conflict tactics in close relationships: A reanlysis of a research instrument. Facet theory: Analysis and design p. 39-46 (1994)
5. Borg, I., Groenen, P.J.: Modern multidimensional scaling: Theory and applications. Springer Science & Business Media (2005)
6. Chapman, B., Jost, G., Pas, R.v.d.: Using OpenMP: Portable Shared Memory Parallel Programming (Scientific and Engineering Computation). The MIT Press (2007)
7. De Leeuw, J.: Applications of convex analysis to multidimensional scaling. Recent Developments in Statistics pp. 133–145 (1977)
8. Dzwinel, W., Blasiak, J.: Method of particles in visual clustering of multidimensional and large data sets. Future Generation Computer Systems **15**(3), 365–379 (1999)

9. Filatovas, E., Podkopaev, D., Kurasova, O.: A visualization technique for accessing solution pool in interactive methods of multiobjective optimization. International Journal of Computers Communications and Control **10**, 508–519 (2015)
10. Garzón, E.M., Moreno, J.J., Martínez, J.A.: An approach to optimise the energy efficiency of iterative computation on integrated GPU–CPU systems. The Journal of Supercomputing **73**(1), 114–125 (2017)
11. Goldberger, J., Gordon, S., Greenspan, H.: An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In: ICCV, pp. 487–493. IEEE Computer Society (2003)
12. Hout, M.C., Goldinger, S.D., Brady, K.J.: MM-MDS: A Multidimensional scaling database with similarity ratings for 240 object categories from the massive memory picture database. PLOS ONE **9**(11), 1–11 (2014)
13. Ingram, S., Munzner, T., Olano, M.: Glimmer: Multilevel MDS on the GPU. IEEE Trans. Vis. Comput. Graph. **15**(2), 249–261 (2009)
14. Kurasova, O., Petkus, T., Filatovas, E.: Visualization of pareto front points when solving multi-objective optimization problems. Information Technology And Control **42**(4), 353–361 (2013)
15. Leng, J., et al.: GPUWattch: Enabling Energy Optimizations in GPGPUs. SIGARCH Comput. Archit. News **41**(3), 487–498 (2013)
16. Medvedev, V., Kurasova, O., Bernatavičienė, J., Treigys, P., Marcinkevičius, V., Dzemyda, G.: A new web-based solution for modelling data mining processes. Simulation Modelling Practice and Theory **76**, 34–46 (2017)
17. Morrison, A., Ross, G., Chalmers, M.: Fast multidimensional scaling through sampling, springs and interpolation. Information Visualization **2**(1), 68–77 (2003)
18. Orts, F., Filatovas, E., Ortega, G., Kurasova, O., Garzón, E.M.: HPC Tool for Multidimensional Scaling. In: Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, vol. 5, pp. 1611–1614. J. Vigo-Aguiar (2017)
19. Orts, F., Filatovas, E., Ortega, G., Kurasova, O., Garzón, E.M.: "Improving the energy efficiency of smacof for multidimensional scaling on modern architectures," *The Journal of Supercomputing*, (2018), doi:10.1007/s11227-018-2285-x.
20. Papenhausen, E., Wang, B., Ha, S., Zelenyuk, A., Imre, D., Mueller, K.: GPU-accelerated incremental correlation clustering of large data with visual feedback. In: Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA, pp. 63–70 (2013)
21. Qiu, J., Bae, S.H.: Performance of windows multicore systems on threading and mpi. Concurrency and Computation: Practice and Experience **24**(1), 14–28 (2012)
22. Shmoys, D.B., Tardos, E.: An approximation algorithm for the generalized assignment problem. Math. Program. **62**(3), 461–474 (1993)
23. Yang, T., Liu, J., McMillan, L., Wang, W.: A fast approximation to multidimensional scaling. In: IEEE workshop on Computation Intensive Methods for Computer Vision (2006)

# Cribado virtual aplicado al potencial electrostático usando un algoritmo evolutivo*

Savíns Puertas-Martín

Departamento de Informática, Universidad de Almería, España
savinspm@ual.es
https://hpca.ual.es/~savins/

**Resumen** Las técnicas de cribado virtual permiten reducir los costes y tiempo necesario para el desarrollo de nuevos fármacos ya que se dedican a buscar en grandes bases de datos aquellos compuestos que más se parecen a un fármaco dado. La medida de similitud entre los distintos compuestos farmacológicos se realiza en base a un descriptor dado. En ese sentido, el descriptor potencial electrostático es uno de los más importantes y de los que mejor caracterizan a los compuestos por sus propiedades intrínsecas. Para encontrar la máxima similitud electrostática entre el fármaco de entrada y un compuesto de la base de datos, la posición relativa entre ambos ha de ser óptima. De este modo se requiere del uso de un algoritmo de optimización como OptiPharm, un algoritmo especialmente diseñado para la búsqueda de compuestos en base a funciones objetivo basadas en la posición 3D de los compuestos. Optipharm es un algoritmo de optimización evolutivo donde una población, formada por un conjunto de poses moleculares, es modificada mediante distintos mecanismos evolutivos con el fin de encontrar soluciones o individuos dentro de la población con la máxima similitud posible. Aparte de los diferentes mecanismos evolutivos, Optipharm incluye conocimiento específico de química computacional para mejorar y acelerar la búsqueda de soluciones eficientes.

* Este trabajo ha sido financiado por el Ministerio de Economía y Competitividad de España (TIN2015-66680-C2-1-R y CTQ2017-87974-R), la Junta de Andalucía (P12-TIC301), la Fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia bajo los proyectos 19419/PI/14 y 18946/JLI/13. Powered@NLHPC: Esta investigación ha sido parcialmente soportada por la infraestructura de supercomputación del NLHPC (ECM-02). Los autores también agradecen los recursos informáticos y el soporte técnico proporcionado por la Plataforma Andaluza de Bioinformática de la Universidad de Málaga. Este trabajo ha sido parcialmente apoyado por las instalaciones informáticas del Centro Extremeño de Tecnologías Avanzadas (CETA-CIEMAT), fundado por el Fondo Europeo de Desarrollo Regional (FEDER). CETA-CIEMAT pertenece a CIEMAT y al Gobierno de España. Savíns Puertas Martín es un beneficiario del programa español "Formación de profesorado universitario", financiado por el Ministerio de Educación, Cultura y Deporte de España. El autor también agradece a Pilar Martínez Ortigosa, Juana López Redondo y Horacio Pérez Sánchez todo el trabajo y apoyo que está recibiendo y que de otra forma este trabajo no sería posible.

## 1. Introducción

El descubrimiento de nuevos fármacos es un proceso muy costoso que requiere aproximadamente de 15 años y aun así, las tasas de éxito son generalmente muy bajas [3]. Para encontrar una solución a esta problemática se han utilizando diferentes enfoques experimentales con el objetivo de encontrar nuevos compuestos farmacológicos que posean las propiedades deseadas. Estos enfoques van desde la medicina tradicional [4] hasta los sistemas de cribado de alto rendimiento (High Throughput Screening) [6]. En relación a éstos últimos, se empezó a utilizar una metodología para reducir el número de compuestos de las bases de datos a subconjuntos mucho más pequeños que podrían caracterizarse experimentalmente. Esta idea se denominó Cribado Virtual (Virtual Screening, VS) y permitió reducir el tiempo y los costes necesarios del proceso de descubrimiento de fármacos [5]. Existen dos métodos de VS dependiendo de la información disponible de los compuestos: los métodos basados en estructuras (Structure Based Virtual Screening, SBVS) y los métodos basados en ligandos (Ligand Based Virtual Screening, LBVS). Los primeros se aplican cuando se conoce la estructura de la proteína objetivo, sin embargo, esta información no está siempre disponible, especialmente cuando el número de estructuras conocidas de la proteína objetivo es muy bajo [9]. En estos casos se usan los métodos LBVS, donde solo se necesitan datos sobre los compuestos conocidos para derivar en otros nuevos y mejorados. Estos datos reciben el nombre de descriptores y cada uno proporciona una información única de los compuestos.

En este trabajo nos centramos en los métodos LBVS utilizando OptiPharm [12] como algoritmo de optimización y realizando el cribado mediante el potencial electrostático. Este descriptor proporciona información de los tipos de átomos que componen cada compuesto en base al potencial electrostático asociado a cada elemento. El potencial electrostático se ha empleado en multitud de trabajos, como es el caso de su uso en soluciones para trastornos hemorrágicos [1] o en la homeostasis de la glucosa [10].

## 2. OptiPharm

En esta sección se describe brevemente OptiPharm. Para conocer en detalle su comportamiento, léase el trabajo original [12].

OptiPharm es un algoritmo evolutivo desarrollado para resolver problemas de compuestos farmacológicos donde la información de sus características se encuentran directamente relacionada con la posición que ocupan en el espacio los átomos de cada compuesto. Esta posición es conocida como pose. Otra ventaja de OptiPharm es que se puede parametrizar en función de diferentes objetivos. Esto significa que puede adaptarse a diferentes tipos de problemas, lo que supone una ventaja en los problemas de cribado virtual en los que los compuestos tienen diferentes tamaños y complejidades.

En la Figura 1 se da una descripción global de OptiPharm. A continuación se describen las diferentes etapas clave del algoritmo:

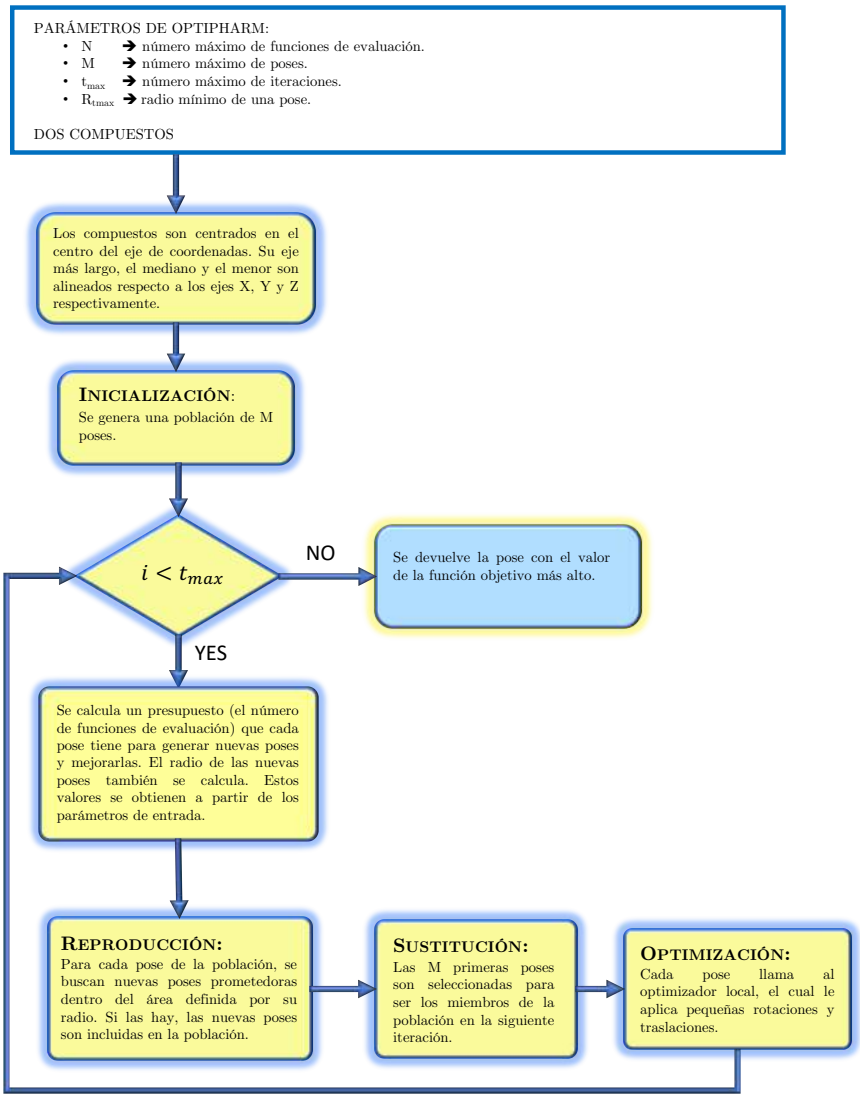



**Figura 1.** Estructura general de OptiPharm.

- *Parámetros de entrada*: Además de los dos compuestos farmacológicos, se deben de indicar otros 4 parámetros adicionales. Éstos establecen la exhaustividad y tiempo de ejecución del algoritmo. Son $N$: número máximo de funciones de evaluación, $M$: número máximo de poses, $t_{max}$: número máximo de iteraciones y $R_{tmax}$: radio de las poses en el último nivel.
- *Inicialización*: Se crea un listado con cinco poses iniciales en el primer nivel. Estas poses cubren todo el espacio de búsqueda pues el radio que define su área de actuación es igual al diámetro del espacio de búsqueda.
- *Reproducción*: Para cada pose en la lista, se crean soluciones candidatas aleatorias en la ventana definida de la pose, y por cada par de soluciones candidatas, se computa el punto medio del segmento que las conecta. Se evalúan tanto el punto medio como los extremos y si alguna de estas soluciones tiene un mejor valor de aptitud que el centro de la pose, entonces ese punto será el nuevo centro, manteniendo el mismo valor de radio. Además, si el valor de la aptitud en un punto medio es peor que en los miembros correspondientes de su par, entonces los miembros del par se insertan en la lista de poses.
- *Sustitución*: Las poses son ordenadas de mayor a menor valor de aptitud y se eliminan las menos aptas hasta que el número de éstas sea inferior o igual al del valor $M$.
- *Optimización*: Se ejecuta el optimizador para cada pose con un número determinado de evaluaciones. Este número se calcula dividiendo el número máximo de evaluaciones asignadas al procedimiento de optimización en el nivel actual por el número máximo de poses en la lista o el tamaño máximo de la población. El optimizador utilizado en este trabajo es el SASS [13].

## 3. Función objetivo: Potencial Electrostático

Esta sección describe el método utilizado para evaluar la similitud del potencial electrostático dado dos compuestos. En la bibliografía existen varios métodos o herramientas que permiten su cálculo, sin embargo, el estado del arte en esta métrica es el Toolkit de ZAP [11] desarrollado por OpenEye (`https://www.eyesopen.com/`). ZAP calcula el potencial electrostático de un compuesto mediante la resolución numérica de la ecuación de Poisson [2], que es la siguiente:

$$\nabla\{\epsilon(r)\nabla\phi(r)\} = -\rho_{mol}(r) \tag{1}$$

donde $\nabla(r)$ es el potencial electrostático, $\epsilon(r)$ es la constante dieléctrica y $\rho_{mol}(r)$ es la distribución de la carga molecular. El potencial electrostático entre dos compuesto se compara por medio de $E_{AB}$:

$$E_{AB} = \int \rho^A(r)\rho^B(r)\Theta^A(r)\Theta^B(r)dr \approx h^3 \sum_{ijk} \phi^A_{ijk}\phi^B_{ijk}\Theta^A_{ijk}\Theta^B_{ijk} \tag{2}$$

donde $\Theta$ es una función que evita considerar el potencial interior de los compuesto como parte de la comparación. La integral que aparece en Eq. 2 es una integral

volumétrica, calculada usando un parámetro $h$ que ajusta la granularidad de la malla de puntos donde va a ser calculado el potencial.

Nótese que la precisión obtenida por la Eq. 2 depende del parámetro $h$ y el número de átomos de los dos compuestos comparados, por tanto, cuanto mayor sea este último número, mayor será el valor de $E_{AB}$ en valor absoluto. Para ser capaz de medir el grado de similitud de dos compuestos, independientemente del número de átomos que los componentes, se utiliza la similaridad de Tanimoto [8] que es calculado de la siguiente forma:

$$Tc = \frac{E_{AB}}{E_{AA} + E_{BB} - E_{AB}} \tag{3}$$

donde $E_{AB}$ es el solapamiento del compuesto A en B. $E_{AA}$ y $E_{BB}$ es el auto-solapamiento de las moléculas A y B respectivamente. La ecuación devuelve un valor en el rango $[-0{,}33, 1]$, donde $-0{,}33$ significa que ambos compuestos tienen la misma carga pero con valor opuesto, 0 significa que no hay solapamiento y 1 significa que las cargas de ambas moléculas son la misma. Esto está explicado en el manual de EON, en la Sección de Teoría (`https://docs.eyesopen.com/eon/theory.html`).

## 4. Resultados

Los estudios computacionales se han realizado utilizando la base de datos de la Agencia Federal del Departamento de Salud y Servicios Humanos de los Estados Unidos (FDA) [14]. La FDA es responsable de proteger y promover la salud pública mediante el control, entre otras cosas, de los medicamentos recetados y de venta libre. Esta agencia proporciona una base de datos que contiene 1751 moléculas, que representa medicamentos aprobados que pueden ser usados con seguridad en humanos en los Estados Unidos. Es una práctica común [7], en el escenario actual, identificar qué pares de compuestos en la base de datos de la FDA comparten un alto grado de similitud de forma.

OptiPharm permite ser parametrizado en base a la calidad y rapidez de los resultados. Para este experimento, se ha configurado para que las ejecuciones sean exhaustivas y robustas. Específicamente, los parámetros han sido: $N = 200000$ evaluaciones de funciones, $M = 5$ poses, $t_{max} = 5$ iteraciones y $R_{tmax} = 1$ como el radio más pequeño posible. Además, dada la naturaleza evolutiva de OptiPharm, los experimentos se han ejecutado 100 veces.

Los experimentos han consistido en seleccionar 10 compuestos *query* aleatorios de la base de datos y compararlos con los 1751 compuestos. En la Tabla 1 se muestra para cada *query* el número de átomos que esta tiene $nAQ$, el compuesto más similar encontrado $BestComp$, su número de átomos $nA$ y por último su valor de similitud de potencial electrostático $Tc$. Téngase en cuenta que el compuesto más similar en todos los casos ha sido el propio compuesto *query* por lo que el compuesto $BestComp$ sería considerado el segundo mejor compuesto encontrado en la base de datos.

Analizando los resultados se puede observar que en todos los casos se obtienen resultados superiores a 0.85, lo que significa que existe al menos 85 % de similitud electrostática y además, OptiPharm se adapta perfectamente a los distintos tamaños de los compuestos ya que encuentra para *querys* de distintos tamaños, compuestos de similar similitud. Esto último es fundamental en los algoritmos aplicados a este tipo de problemas pues la variedad de compuestos en una misma base de datos puede tener una diferencia de 200 átomos entre el compuesto con menor número de átomos y el de mayor número.

| query | nAQ | BestComp | nA | Tc |
|---|---|---|---|---|
| DB01352 | 29 | DB00418 | 35 | 0.96 |
| DB01365 | 30 | DB01191 | 33 | 0.96 |
| DB06216 | 37 | DB01158 | 30 | 0.94 |
| DB07615 | 40 | DB04552 | 28 | 0.88 |
| DB00246 | 50 | DB09224 | 42 | 0.87 |
| DB09237 | 54 | DB01209 | 42 | 0.88 |
| DB01621 | 66 | DB08810 | 60 | 0.85 |
| DB08903 | 69 | DB01242 | 46 | 0.87 |
| DB01419 | 70 | DB01337 | 101 | 0.91 |
| DB04786 | 120 | DB09267 | 31 | 0.89 |

**Tabla 1.** Resultados obtenidos para 10 compuestos de la base de datos FDA. Para cada *query*, se muestra su número de átomos $nAQ$, el compuesto más parecido electrostáticamente $BestComp$, el número de átomos de este compuesto $nA$ y el valor de similitud $Tc$.

## 5. Conclusiones

En este trabajo se ha presentado un algoritmo evolutivo llamado OptiPharm y que se ha aplicado a LBVS utilizando el potencial electrostático. Utilizando la base de datos de la FDA se ha demostrado que es capaz de encontrar soluciones de alta calidad independientemente del número de átomos de los compuestos, lo que es fundamental en este tipo de problemas.

Como trabajo futuro, y en base a que OptiPharm es intrínsecamente paralelo pues cada pose en el listado genera nuevos candidatos de manera independiente al resto de poses de la lista, se diseñarán e implementarán diferentes paradigmas de programación basados en arquitecturas de memoria compartida y distribuida.

## Referencias

1. Boström, J., Grant, J.A., Fjellström, O., Thelin, A., Gustafsson, D.: Potent Fibrinolysis Inhibitor Discovered by Shape and Electrostatic Complementarity to the Drug Tranexamic Acid. Journal of Medicinal Chemistry **56**(8), 3273–3280 (2013)

2. Böttcher, C., Belle, O.V., Belle, B.: Theory of electric polarization (1974)
3. Drews, J.: Drug discovery: a historical perspective. Science **287**(5460), 1960–1964 (2000)
4. Fu, X., Mervin, L.H., Li, X., Yu, H., Li, J., Mohamad Zobir, S.Z., Zoufir, A., Zhou, Y., Song, Y., Wang, Z., Bender, A.: Toward understanding the cold, hot, and neutral nature of chinese medicines using in silico mode-of-action analysis. Journal of Chemical Information and Modeling **57**(3), 468–483 (2017)
5. Geppert, H., Vogt, M., Bajorath, J.: Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. Journal of chemical information and modeling **50**(2), 205–216 (2010)
6. Glick, M., Jenkins, J.L., Nettles, J.H., Hitchings, H., Davies, J.W.: Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. Journal of chemical information and modeling **46**(1), 193–200 (2006)
7. den Haan, H., Morante, J.J.H., Perez-Sanchez, H.: Computational evidence of a compound with nicotinic $\alpha 4\beta 2$-ach receptor partial agonist properties as possible coadjuvant for the treatment of obesity. bioRxiv (2016)
8. Jaccard, P.: Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 241–272 (1901)
9. Lipinski, C.A.: Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. Advanced drug delivery reviews **101**, 34–41 (2016)
10. Markt, P., Petersen, R.K., Flindt, E.N., Kristiansen, K., Kirchmair, J., Spitzer, G., Distinto, S., Schuster, D., Wolber, G., Laggner, C., Langer, T.: Discovery of novel ppar ligands by a virtual screening approach based on pharmacophore modeling, 3d shape, and electrostatic similarity screening. Journal of Medicinal Chemistry **51**(20), 6303–6317 (2008)
11. OpenEye Scientific Software: OEChem Toolkit, `https://www.eyesopen.com/zap-tk`
12. Puertas-Martín, S., Redondo, J.L., Ortigosa, P.M., Pérez-Sánchez, H.: Optipharm: An evolutionary algorithm to compare shape similarity. Scientific Reports **9**(1) (2019)
13. Solis, F.J., Wets, R.J.B.: Minimization by Random Search Techniques. Mathematics of Operations Research **6**(1), 19–30 (1981)
14. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J.: Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research **34**, D668–D672 (2006)

# Gestión de recursos heterogéneos en «energy hubs» con autoconsumo

Jerónimo Ramos Teodoro

Centro mixto CIESOL, Campus de Excelencia Internacional Agroalimentario (ceiA3), Universidad de Almería, Ctra. Sacramento s/n, Almería 04120, España
{jeronimo.rt}@ual.es

**Resumen** El concepto de «energy hub» y su metodología de modelado son herramientas útiles para resolver problemas de reparto de recursos. Este documento recoge las aportacioens realizadas durante el año 2018 representar con mayor precisión las plantas reales y reducir la cantidad de variables de decisión. La principal innovación es considerar dispositivos que consumen un recurso que no está relacionado con la cantidad de salida producida, incluyendo variables de decisión binarias en ciertas salidas del «energy hub». En segundo lugar, se define un vector de ruta para tener en cuenta los flujos de recursos dentro del sistema en lugar de emplear una variable para cada rama entre los componentes. La tercera innovación consiste en un vector adicional para expresar la cantidad de recursos de salida vendidos dal exterior, incorporando restricciones para aquellos recursos que se exportan e importan a través del mismo medio. Todo esto se incluye en un modelo generalizado y luego se aplica a un ejemplo de planta real, que incluye recursos múltiples y heterogéneos. Los resultados comparativos entre días con diferentes demandas, condiciones climáticas y precios de la electricidad validan el enfoque propuesto y los beneficios del uso de fuentes renovables.

## 1. Introducción

En los últimos años, las políticas energéticas destinadas a aumentar la eficiencia en los procesos de producción, transporte, consumo y almacenamiento han conducido a enfoques basados en la descentralización de estos procesos y la combinación de diferentes tipos de energía para beneficiarse de la sinergia derivada del uso de los recursos e infraestructuras locales disponibles. Sin embargo, debido a su naturaleza intermitente, en muchos casos se requieren sistemas de almacenamiento y estrategias de gestión que desvinculen la generación de la demanda para ser económicamente viables [1].

Conceptos recientes como la multigeneración distribuida (DMG) [2] y los sistemas multi-energía (MES) [3] han llegado a establecer un marco de investigación general para sistemas que incluyen varios vectores energéticos, integrándolos en programas de respuesta a la demanda [4]. Dentro del llamado MES, el enfoque «energy hub» se usa ampliamente como un modelo simplificado de las interacciones dentro de sistemas de diversa complejidad atendiendo a su estructura de

entrada-salida. Una definición formal del concepto «energy hub», o concentrador de energía, fue dada por primera vez por Geidl et al. [5] en 2007: «una unidad donde se pueden convertir, acondicionar y almacenar múltiples portadoras de energía».

Desde entonces, muchos autores han aplicado este concepto a diferentes casos, como la gestión de recursos, la introducción de algoritmos de optimización robustos [6], metodologías para mejorar la representación de restricciones operacionales [7] o para facilitar el modelado automático de configuraciones arbitrarias de «energy hubs» [8]. El concepto también se usa en problemas relacionados con el despacho económico [9] o la configuración óptima de dispositivos [10]. Una revisión reciente sobre los recursos y convertidores comunes en «energy hubs» [11] muestra que la mayoría de los trabajos se centran en las tecnologías de calefacción y refrigeración térmica, prestando menos atención a otros recursos materiales (como combustibles sólidos, agua, dióxido de carbono o hidrógeno). Esto los hace interesantes desde el punto de vista de la investigación. Dado que los modelos de concentrador de energía pueden usarse tanto para flujos de energía como de materiales, el término «recurso» es preferible para referirse a ellos.

Un obstáculo existente es que los marcos legales para las redes MES aún están en desarrollo y cada red de suministro y distribución de recursos generalmente opera bajo sus propias reglas. Los sistemas de energía eléctrica tienen uno de los marcos legales más extensos, que ha establecido las bases fundamentales para el desarrollo de redes inteligentes [12]. Esto ha sido gracias a los recientes avances en las tecnologías de la información y la comunicación, así como en la automatización y la electrónica. Por ejemplo, en el contexto español, algunos trabajos ya se han ocupado de la gestión de la microrred y su integración en el mercado de la electricidad como agente del sistema (tanto como consumidor directo como productor) [13]. Respecto a los consumidores indirectos, que no se consideran agentes del mercado [14] y, por lo tanto, no participan en el mercado de la electricidad, las leyes que regulan el autoconsumo [15,16] establecen dos tipos de autoconsumidores (según los las características de la instalación y el titular del contrato), como se describe a continuación.

Por lo tanto, a pesar de las diferencias legales entre países, el autoconsumo y los sistemas basados en renovables se convierten en una consideración interesante para las pequeñas industrias [17], comercioes [18] o edificios residenciales [19], debido a sus potenciales beneficios económicos y ambientales [20]. Este es el ámbito en el que se desarrollan los proyecto ENERPRO [21] y CHROMAE [22], dentro de los cuales se lleva a cabo el trabajo de tesis del que deriva este documento, bajo la concesión de una beca FPI.

### 1.1. Objetivos de la tesis

En los proyectos mencionados anteriormente se abarca el análisis, diseño y aplicación de técnicas de modelado, predicción, estimación, control y optimización para conseguir una gestión óptima de energía y recursos en entornos pro-

ductivos que hacen uso de fuentes renovables y de sistemas de almacenamiento, especialmente distritos agroindustriales.

El principal objetivo que persigue esta tesis doctoral es el diseño de técnicas de gestión óptima de recursos en estos entornos, para lo cual se requieren modelos estáticos y dinámicos con distintos niveles de abstracción y la implementación de técnicas de control y optimización (incluyendo el control predictivo basado en modelo) para conseguir adecuar la producción a la demanda. Por tanto, se determinan los siguientes objetivos particulares:

1. Desarrollo de modelos físicos de sistemas generadores, consumidores y almacenadores de recursos con una filosofía modular y distintos niveles de abstracción que permitan la simulación dinámica a corto, medio y largo plazo a nivel de sistema y su uso en la optimización. Se hará especial énfasis en paradigmas de modelado tipo energy hub.
2. Caracterización de la generación, la demanda y las perturbaciones. Diseño de predictores y estimadores.
3. Validación experimental de los modelos desarrollados en condiciones estacionarias y dinámicas.
4. Aplicación de los modelos diseñados en el desarrollo de estrategias de optimización y coordinación de la producción desde los puntos de vista energético, económico y de seguridad.
5. Desarrollo de estudios de análisis de sensibilidad e incertidumbre usando indicadores basados en la energía y en aspectos económicos en la instalación productiva de referencia. Desarrollo de estrategias de control robustos.
6. Aplicación a las instalaciones del proyecto ENERPRO descritas en el siguiente epígrafe.

### 1.2. Breve descripción de la planta de ensayos de ENERPRO

A modo de planta de pruebas se dispone de un sistema formado por el Centro de Investigación en Energía Solar (CIESOL), un invernadero, un parking fotovoltaico y una desaladora solar que demandan y producen diferentes recursos conforme representa la figura 1.

Por un lado, CIESOL provee de energía eléctrica fotovoltaica al resto de elementos y de energía térmica mediante el uso de captadores solares. Por otro, tato en la caldera del invernadero como en la planta desaladora se produce calor a partir de energía renovable y agua potable, en esta última, para abastecer al resto del sistema. Aunque se trata de un conjunto autosostenible, cuenta con suministro hídrico y eléctrico desde la red para ejercer de apoyo. En los siguientes párrafos se detallan los elementos de interés para el propósito de este trabajo, partiendo de la recopilación realizada en trabajos previos:

- El sistema de producción de CIESOL se compone de 42 módulos fotovoltaicos Atersa A-222P, 80 captadores planos Solaris CP1, una máquina de absorción Yazaki WFC SC20 y una bomba de calor reversible Ciatesa Hidripack WE 360. Una descripción detallada del edificio puede encontrarse en [23].

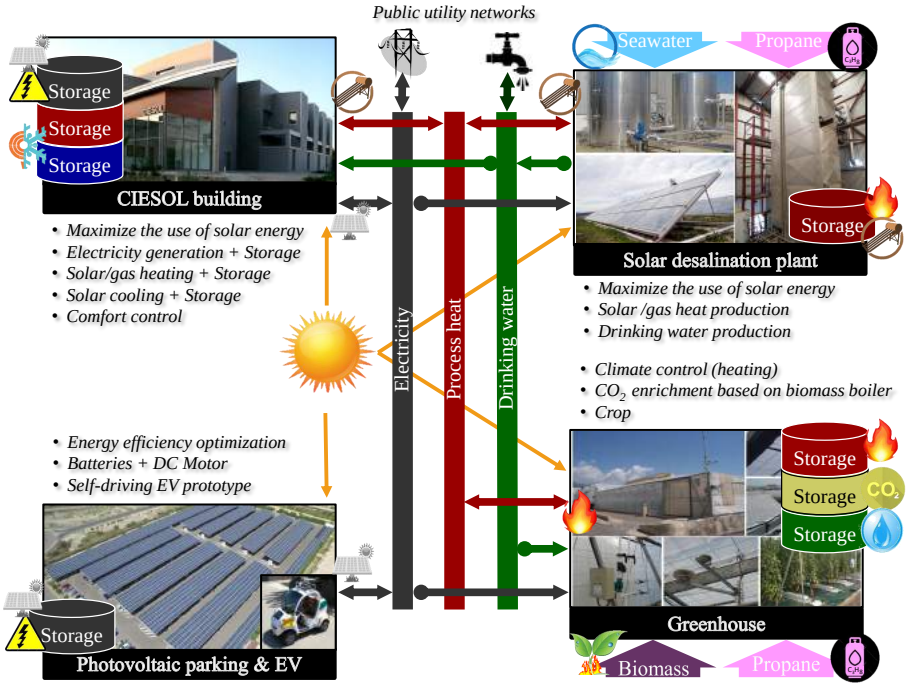



**Figura 1.** Enfoque y diagrama funcional de la planta de pruebas

- El sistema AQUASOL de la Plataforma Solar de Almería [24] consiste en una planta desaladora multiefecto (MED) manufacturada por ENTROPIE, que funciona a partir de un campo de 252 captadores CPC Ao Sol 1.12x y el apoyo de una caldera ATTSU RL200 de gas propano.
- El invernadero del proyecto ENERPRO [25] está instalado en la Estación Experimental de la Fundación Cajamar (El Ejido, Almería). Se trata de un invernadero tipo parral con una superficie de 877 m2 cuyo sistema de calefacción está constituido por un calefactor GP 95 de propano y por una caldera Missouri 150 000 de biomasa.
- El parking fotovoltaico de la Universidad de Almería cuenta con 483 paneles CONERGY PA 264P, 24 paneles CONERGY POWER PLUS 240M y 72 paneles FIRST SOLAR FS-380, con una potencia pico total de la instalación de 1176,48 kW y una potencia nominal de 1015 kW.

## 2. Avances y desarrollo de la tesis

Actualmente se dispone de un modelo global de la planta productiva (véase la figura 2 y el cuadro 1), que se emplea en algoritmos de optimización basados en programación lineal en enteros mixta («mixed integer linear programming»,

abreviada usalmente como MILP) para determinar el reparto económico de recursos en diferentes escenarios.

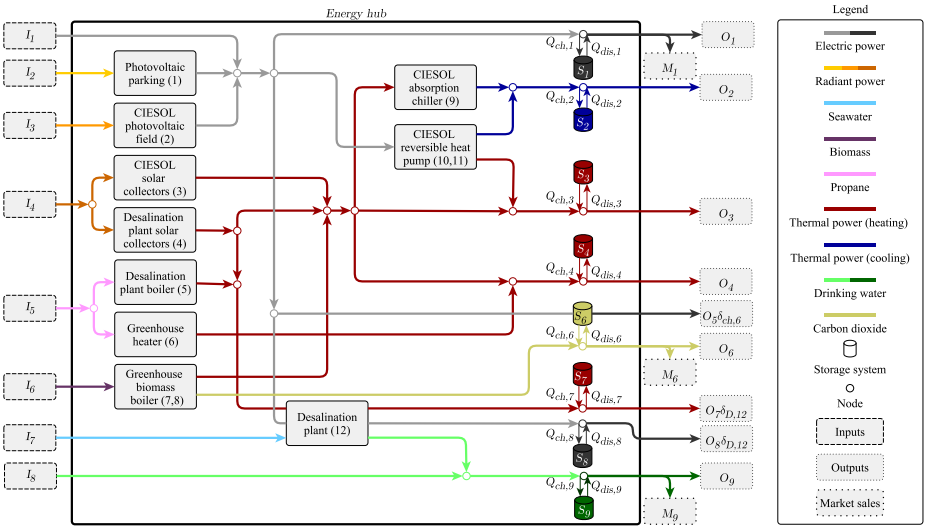



**Figura 2.** Modelo energy hub propuesto para la gestión de la planta ENERPRO

Algunos elementos de la planta, como las instalaciones de energía solar, cuentan con modelos en detalle mientras que para el resto se emplean modelos simplificados o basados en datos (autoregresivos y series temporales), como es el caso de las previsiones para la demanda. Se han llevado a cabo simulaciones de la planta en diversos escenarios, considerando predicciones deterministas, dando lugar a diferentes publicaciones [26,27,28,29,30,31]. Además, durante la estancia realizada entre julio y septiembre de 2018 en la Universidad Federal de Santa Catarina (Florianópolis, Brasil), se han realizado avances en la formulación de estrategias de control bajo incertidumbre y el empleo de modelos de predicción de la radiación solar basados en redes neuronales.

A lo largo del año 2018 pueden destacrase las siguientes contribuciones [29]:

1. Complementar modelos anteriores para incluir la posibilidad de representar con mayor precisión ciertos procesos, como vender recursos de salida y agregar cargas relacionadas con el estado operativo (encendido o apagado) de ciertos dispositivos, lo que garantiza resultados más económicos; y para reducir el número de variables de decisión para modelar centros de energía complejos, lo que implica reducir el esfuerzo de cálculo.
2. Ejemplo de modelado con múltiples recursos materiales y energéticos, incluyendo entradas inusuales (como el agua de mar y la biomasa), salidas ($CO_2$ enriquecimiento para un invernadero [25]) y convertidores (una planta de desalinización por energía solar y una máquina de absorción).

**Cuadro 1.** Descrición de entradas, salidas y venta de recursos

| Variable | Descripción | Unidades |
|---|---|---|
| $I_1$ | Electricidad de la red pública eléctrica | kW |
| $I_2$ | Radiación solar incidente en los módulos fotovoltaicos del parking | kW |
| $I_3$ | Radiación solar incidente en los módulos fotovoltaicos de CIESOL | kW |
| $I_4$ | Radiación solar incidente en los captadores solares | kW |
| $I_5$ | Propano para los sistemas de combustión | kg/h |
| $I_6$ | Pélets de madera para la caldera de biomasa | kg/h |
| $I_7$ | Agua de mar para la planta desaladora | $m^3/h$ |
| $I_8$ | Agua potable de la red de abastecimiento pública | $m^3/h$ |
| $O_1$ | Electricidad para CIESOL y el invernadero | kW |
| $O_2$ | Potencia térmica (refrigeración) demandada por CIESOL | kW |
| $O_3$ | Potencia térmica (calefacción) demandada por CIESOL | kW |
| $O_4$ | Potencia térmica (calefacción) demandada por el invernadero | kW |
| $O_5$ | Electricidad para la bomba de $CO_2$ del invernadero | kW |
| $O_6$ | $CO_2$ para el invernadero | kg/h |
| $O_7$ | Potencia térmica (calefacción) demandada por la desaladora | kW |
| $O_8$ | Electricidad para la desaladora | kW |
| $O_9$ | Agua para CIESOl y el invernadero | $m^3/h$ |
| $M_1$ | Electricidad vendida a través de la red pública eléctrica | kW |
| $M_6$ | $CO_2$ descargado a la atmósfera desde el almacenamiento | kg/h |
| $M_9$ | Agua vendida a través de la red de abastecimiento pública | $m^3/h$ |

3. Probar la validez del enfoque propuesto mediante la simulación de dos casos diferentes en sistemas con autoconsumo, en los que la programación operativa se determina considerando las variaciones en el precio de la electricidad a lo largo del día.

Dicho modelo establece las ecuaciones matemáticas entre las entradas y las salidas consideradas que permiten determinar el reparto óptimo de recursos, satisfaciendo las necesidades de demanda. Para cada uno de los elementos que conforma el energy hub es posible sustituir los modelos ya disponibles por otros, de mayor o menor complejidad, que se desarrollen en el futuro. El sistema completo se ha implementado en el entorno de programación MATLAB. En [29] se presentan los resultados para un día típico de invierno (31/01/2014) y otro de verano (01/08/2014).

En los escenarios propuestos recientemente se considera la posibilidad de operar en el mercado eléctrico como productor y consumidor directo, lo cual supone un enfoque diferente a los anteriores desde el punto de vista económico. El conjunto se puede clasificar como de autoconsumo tipo 2 y, por lo tanto, el excedente de energía producido en las instalaciones fotovoltaicas se puede vender a través de la red pública. Así pues, el precio de la electricidad juega un papel fundamental en el problema de optimización, ya que delimita el tamaño del horizonte considerado y determina los elementos de la función de coste. Los precios horarios de la electricidad se obtienen desde la web de OMIE [14] y se actualizan según los horarios de publicación establecidos por la compañía.

Debido a esta logística, se emplea un horizonte deslizante variable:

- De 0 h a 18 h (UTC + 1 / UTC + 2, según el horario de invierno o verano), la programación de la operación se calcula acortando $H$ para que sea igual a la diferencia entre la hora actual y la medianoche, el período cuando el precio de la electricidad es conocida.
- Desde las 18 h hasta las 24 h (UTC + 1 / UTC + 2, según el horario de invierno o verano), el precio de la electricidad se publica durante todo el día siguiente, por lo que $H$ toma un valor de 24 h.

En ambos casos, los elementos de la función de coste se actualizan de acuerdo con los precios publicados por OMIE con respecto a los mercados diarios e intradiarios, dejando una hora de margen desde el momento en que son publicados.

Los principales resultados se resumen en la Tabla 2, en términos de energía, masa o volumen demandados o suministrados para los vectores $\boldsymbol{I}$, $\boldsymbol{O}$ y $\boldsymbol{M}$, durante los días simulados. Además, se incluye el costo total de importación (valores positivos) o venta (valores negativos) de recursos. Aunque la demanda difiere entre el verano y el día de invierno, especialmente en relación con la energía térmica, en ambos casos el costo total de operación es bastante similar: menos de una diferencia de 3 €. Las mayores demandas de 01/08/2014 fueron compensadas por las ventas de electricidad y agua a través de las redes de servicios públicos.

## 3. Conclusiones

Una propuesta integral para la gestión de recursos, como el concepto de concentrador de energía presentado en este documento, puede ser un instrumento adecuado para los problemas de planificación de recursos. Los resultados obtenidos demuestran que el enfoque adoptado es apropiado incluso cuando la programación depende de decisiones mutuamente excluyentes (por ejemplo, resolver cuando un dispositivo debe activarse o no), así como la validez de los nuevos elementos introducidos en el marco de los «energy hub». La programación se ha realizado cuidadosamente de manera que no solo se pueden considerar diferentes escenarios sino también diferentes centros de energía, simplemente definiendo sus estructuras (entradas, salidas, dispositivos y conexiones) de una manera directa.

El enfoque de gestión propuesto ayuda a aumentar la precisión en la representación del proceso sin aumentar la complejidad computacional. Esto se logra definiendo el vector de ruta en lugar de incluir una variable de decisión para cada rama [8] entre dispositivos o entre ellos y los nodos de entrada o salida. Los tiempos de cómputo requeridos para resolver el problema de programación permiten aumentar la complejidad del modelo y aún así hacerlo aplicable en tiempo real. La elección de la longitud del horizonte deslizante, por ejemplo, es uno de los principales factores determinantes del tiempo de cálculo y su aumento podría mejorar los resultados obtenidos. Sin embargo, esto requiere el desarrollo de estimadores del precio de la electricidad además de los modelos de demanda.

**Cuadro 2.** Input, output and market total demand/supply and costs

| | Summer day | | Winter day | |
|---|---|---|---|---|
| Variable | Accumulated | Cost | Accumulated | Cost |
| $I_1$ | 1150.28 kWh | 116.81 € | 1171.32 kWh | 141.82 € |
| $I_2$ | 5102.92 kWh | 60.73 € | 2615.97 kWh | 22.89 € |
| $I_3$ | 39.91 kWh | 0.20 € | 262.63 kWh | 2.18 € |
| $I_4$ | 3189.96 kWh | 0 € | 606.42 kWh | 0 € |
| $I_5$ | 0.19 kg | 0.33 € | 0 kg | 0 € |
| $I_6$ | 15 kg | 3.83 € | 15 kg | 3.83 € |
| $I_7$ | 51 $m^3$ | 0 € | 0 $m^3$ | 0 € |
| $I_8$ | 0.97 $m^3$ | 0.53 € | 3.42 $m^3$ | 1.87 € |
| $O_1$ | 1611.58 kWh | - | 1530.25 kWh | - |
| $O_2$ | 386.96 kWh | - | 5.37 kWh | - |
| $O_3$ | 0 kWh | - | 182.24 kWh | - |
| $O_4$ | 37.93 kWh | - | 13.48 kWh | - |
| $O_5$ | 2.8 kWh | - | 2.8 kWh | - |
| $O_6$ | 20.74 kg | - | 15.52 kg | - |
| $O_7$ | 927.00 kWh | - | 0 kWh | - |
| $O_8$ | 68.4 kWh | - | 0 kWh | - |
| $O_9$ | 2.02 $m^3$ | - | 3.37 $m^3$ | - |
| $M_1$ | 41.32 kWh | -2.80 € | 1.81 kWh | -0.11 € |
| $M_6$ | 5.65 kg | 0 € | 10.88 kg | 0 € |
| $M_9$ | 9.82 $m^3$ | -4.39 € | 0 $m^3$ | 0 € |
| Total | - | 175.23 € | - | 172.48 € |

En términos de trabajo futuro, se realizarán estudios más profundos para mejorar la precisión de algunos parámetros y para lidiar con su incertidumbre (perfiles de demanda, factores de conversión, precios de recursos, etc.), así como para desarrollar los modelos y estimadores mencionados anteriormente. Otra de las líneas actuales de investigación, es la interacción entre múltiples concentradores de energía. Además, al hacer que las variables binarias correspondan a las entradas y los valores enteros del dispositivo, junto con una selección adecuada de los límites superior e inferior, los dispositivos que operan se podría representar una carga parcial; por lo tanto, realizar una comparación con una aproximación por etapas [7] podría resultar un análisis interesante. Otro problema es la estrategia de gestión empleada, ya que se formula de acuerdo con el marco legal de autoconsumo. Las alternativas basadas en la inclusión del centro de energía como un agente del mercado de la electricidad podrían evaluarse para descubrir el escenario más económico. Sin embargo, el enfoque actual sigue siendo viable para un gran número de prosumidores que podrían beneficiarse del autoconsumo.

## Referencias

1. Korkas, C.D., Baldi, S., Michailidis, I., Kosmatopoulos, E.B.: Occupancy-based demand response and thermal comfort optimization in microgrids with renewable energy sources and energy storage. Appl. Energy **163** (Feb. 2016) 93–104

2. Chicco, G., Mancarella, P.: Distributed multi-generation: A comprehensive view. Renew. Sustain. Energy Rev. **13**(3) (Apr. 2009) 535–551
3. Mancarella, P.: MES (multi-energy systems): An overview of concepts and evaluation models. Energy **65** (Feb. 2014) 1–17
4. Wang, J., Zhong, H., Ma, Z., Xia, Q., Kang, C.: Review and prospect of integrated demand response in the multi-energy system. Appl. Energy **202** (Sep. 2017) 772–782
5. Geidl, M., Koeppel, G., Favre-Perrod, P., Klöckl, B., Andersson, G., Fröhlich, K.: Energy hubs for the future. IEEE Power Energy Mag. **5**(1) (Jan.–Feb. 2007) 24–30
6. Parisio, A., Del Vecchio, C., Vaccaro, A.: A robust optimization approach to energy hub management. Int. J. Electr. Power Energy Syst. **42**(1) (Nov. 2012) 98–104
7. Evins, R., Orehounig, K., Dorer, V., Carmeliet, J.: New formulations of the 'energy hub' model to address operational constraints. Energy **73** (Aug. 2014) 387–398
8. Wang, Y., Cheng, J., Zhang, N., Kang, C.: Automatic and linearized modeling of energy hub and its flexibility analysis. Appl. Energy **211** (Feb. 2018) 705–714
9. Beigvand, S.D., Abdi, H., La Scala, M.: A general model for energy hub economic dispatch. Appl. Energy **190** (Mar. 2017) 1090–1111
10. Wang, Y., Zhang, N., Zhuo, Z., Kang, C., Kirschen, D.: Mixed-integer linear programming-based optimal configuration planning for energy hub: Starting from scratch. Appl. Energy **210** (Jan. 2018) 1141–1150
11. Mohammadi, M., Noorollahi, Y., Mohammadi-Ivatloo, B., Yousefi, H.: Energy hub: From a model to a concept – A review. Renew. Sustain. Energy Rev. **80** (Dec. 2017) 1512–1527
12. Vasconcelos, J.: Survey of regulatory and technological developments concerning smart metering in the European Union electricity market. European Union Institute - Robert Schuman Centre for Advanced Studies (Policy Papers 2008/01) (Sep. 2008)
13. Bordons, C., García-Torres, F., Valverde, L.: Optimal energy management for renewable energy microgrids. Rev. Iberoam. Automática e Informática Ind. RIAI **12**(2) (Apr.—Jun. 2015) 117–132
14. OMI-Polo Español S.A. (OMIE): OMIE web site. [En línea]. Disponible en: http://www.omie.es/en/inicio Accedido: 21/01/2019.
15. Ministerio de Industria, Energía y Tourismo: Real Decreto 900/2015, de 9 de octubre, por el que se regulan las condiciones administrativas, técnicas y económicas de las modalidades de suministro de energía eléctrica con autoconsumo y de producción con autoconsumo. [En línea]. Disponible en: https://www.boe.es/boe/dias/2015/10/10/pdfs/BOE-A-2015-10927.pdf Accedido: 21/01/2019.
16. Jefatura del Estado: Real Decreto-ley 15/2018, de 5 de octubre, de medidas urgentes para la transición energética y la protección de los consumidores. [En línea]. Disponible en: https://www.boe.es/boe/dias/2018/10/06/pdfs/BOE-A-2018-13593.pdf Accedido: 21/01/2019.
17. Elsner, W., Wysocki, M., Niegodajew, P., Borecki, R.: Experimental and economic study of small-scale chp installation equipped with downdraft gasifier and internal combustion engine. Appl. Energy **202** (Sep. 2017) 213–227
18. Merei, G., Moshövel, J., Magnor, D., Sauer, D.U.: Optimization of self-consumption and techno-economic analysis of pv-battery systems in commercial applications. Appl. Energy **168** (Apr. 2016) 171–178
19. Luthander, R., Widén, J., Nilsson, D., Palm, J.: Photovoltaic self-consumption in buildings: A review. Appl. Energy **142** (Mar. 2015) 80–94

20. Bertsch, V., Geldermann, J., Lühn, T.: What drives the profitability of household pv investments, self-consumption and self-sufficiency? Appl. Energy **204** (Oct. 2017) 1–15
21. Grupo de Investigación Automática, Robótica y Mecatrónica (ARM-TEP197): Proyecto ENERPRO. [En línea]. Disponible en: http://www2.ual.es/enerpro/ Accedido: 21/01/2019.
22. Grupo de Investigación Automática, Robótica y Mecatrónica (ARM-TEP197): Proyecto CHROMAE. [En línea]. Disponible en: http://www2.ual.es/chromae/ Accedido: 21/01/2019.
23. Castilla, M.M., Álvarez, J.D., Rodríguez, F., Berenguel, M.: Comfort Control in Buildings. Springer (Jun. 2014)
24. Alarcón Padilla, D.C., Blanco Gálvez, J., García Rodríguez, L., Gernjak, W., Malato Rodríguez, S.: First experimental results of a new hybrid solar/gas multi-effect distillation system: the AQUASOL project. Desalination **220**(1-3) (Mar. 2008) 619–625
25. Sánchez-Molina, J.A., Reinoso, J.V., Acién, F.G., Rodríguez, F., López, J.C.: Development of a biomass-based system for nocturnal temperature and diurnal $CO_2$ concentration control in greenhouses. Biomass and Bioenergy **67** (Aug. 2014) 60–71
26. Ramos-Teodoro, J.: Gestión energética de un sistema de producción heterogéneo bajo el paradigma energy hub. Trabajo Fin de Máster, Universidad Carlos III de Madrid (2017)
27. Ramos-Teodoro, J., Álvarez, J.D., Rodríguez, F., Berenguel, M.: Gestión económica de energy hubs con recursos heterogéneos mediante MINLP. In: IV Simposio CEA de Modelado, Simulación y Optimización, Universidad de Valladolid (2018)
28. Ramos-Teodoro, J., Rodríguez, F., Berenguel, M.: Modelado basado en el paradigma de los energy hubs de una explotación agraria bajo invernadero con apoyo de energías renovables. In: I Symposium Ibérico de Ingeniería Hortícola, Universidad de Santiago de Compostela - Campus Terra (2018)
29. Ramos-Teodoro, J., Rodríguez, F., Berenguel, M., Torres, J.L.: Heterogeneous resource management in energy hubs with self-consumption: Contributions and application example. Appl. Energy **229** (Nov. 2018) 537–550
30. Ramos-Teodoro, J., Rodríguez, F., Berenguel, M.: Modelado de instalaciones fotovoltaicas para la gestión de un energy hub con recursos heterogéneos. In: XVI Simposio CEA de Ingeniería de Control, Universidad de Almería (2018)
31. Ramos-Teodoro, J., Rodríguez, F., Berenguel, M.: Estudio comparativo de gestión energética en una planta agroindustrial con autoconsumo. In: Congreso de Jóvenes Investigadores en Ciencias Agroalimentarias, Centro de Investigación en Agrosistemas Intensivos Mediterráneos y Biotecnología Agroalimentaria de la Universidad de Almería (CIAIMBITAL) (2018)

# Modelado y Optimización de Problemas en Sanidad vía Computación de Altas Prestaciones.

## Calibrado de parámetros en modelos epidemiológicos complejos: una aproximación multi-objetivo.

Miriam R. Ferrández

Dept. de Informática, Universidad de Almería, ceiA3, Ctra. Sacramento, La Cañada de San Urbano, 04120 Almería, España
mrferrandez@ual.es

**Abstract.** Los modelos epidemiológicos permiten predecir la propagación de las enfermedades, sin embargo, debe realizarse un trabajo previo de calibrado de algunos de los parámetros involucrados. En este trabajo, proponemos una metodología novedosa para ajustar esos parámetros. Se basa en resolver un problema de optimización multi-objetivo, cuyas funciones objetivo miden la precisión del modelo. En concreto, hemos considerado el modelo denominado Between-Countries Disease Spread (Be-CoDiS), que describe la propagación en un conjunto de países teniendo en cuenta los movimientos migratorios entre ellos. Como resultado, utilizando algunos datos reales del brote de Ébola, como el número de casos detectados y el número de muertes, hemos probado que la metodología propuesta es capaz de encontrar un conjunto de valores para los parámetros de modo que el modelo ajusta con precisión la propagación de la epidemia en un conjunto amplio de países.

## 1 Introducción

En 2014, el brote del virus del Ébola llevó a una seria preocupación sobre la capacidad de las autoridades para predecir y controlar las enfermedades epidémicas y su propagación entre países. En este contexto, sugieron algunas aproximaciones novedosas para modelar estas situaciones, como el modelo denominado *Between-Countries Disease Spread* (Be-CoDiS) propuesto en [1]. A pesar de las precisas predicciones logradas con este modelo, se evidenció el reto que supone el ajuste de los valores para los parámetros epidemiológicos involucrados. Estos parámetros dependen de las características de cada país como, por ejemplo, su desarrollo económico y su demografía. Además, incluso pueden depender del tiempo.

Este trabajo trata de conseguir una metodología de ajuste global, en la que se considera un conjunto de países conectados por sus movimientos migratorios. Para ello, se emplea el modelo Be-CoDiS y se define un problema multi-objetivo. Cuando la propagación de la epidemia se simula numéricamente de acuerdo con este modelo, éste devuelve la evolución de la infección. Como queremos que esta evolución sea lo más fiel posible a la realidad, las funciones objetivo son

formuladas como las diferencias entre esos datos predichos por el modelo y los reales.

## 2 El modelo epidemiológico: Be-CoDiS

El modelo determinístico conocido por sus siglas Be-CoDiS describe la evolución de una epidemia en un grupo de $N_{\text{co}} \in \mathbb{N}$ países, teniendo en cuenta los flujos migratorios entre ellos. Es un modelo basado en compartimentos, de modo que los individuos de la población de cada país $i \in \{1, \ldots, N_{\text{co}}\}$ se clasifican en los siguientes estados disjuntos: susceptible, infectado, infeccioso, hospitalizado, recuperado, fallecido y enterrado. El número de personas en el país $i$ en el instante $t$ pertenecientes a cada uno de esos estados se denotan, respectivamente, por $S(i,t), E(i,t), I(i,t), H(i,t), R(i,t), D(i,t)$ y $B(i,t)$. Considerando $NP(i,t)$ el número total de personas en el país $i$ en el instante $t$, entonces $NP(i,t) = S(i,t) + E(i,t) + I(i,t) + H(i,t) + R(i,t)$.

Este modelo Be-CoDiS, que fue propuesto y validado en [1], viene dado por las siguientes ecuaciones:

$$
\begin{aligned}
\frac{\mathrm{d}S(i,t)}{\mathrm{d}t} =& -\frac{S(i,t)\Big(m_I(i,t)\beta_I(i)I(i,t) + m_H(i,t)\beta_H(i)H(i,t) + m_D(i,t)\beta_D(i)D(i,t)\Big)}{\mathrm{NP}(i,t)} \\
& - \mu_m(i)S(i,t) + \mu_n(i)\Big(S(i,t) + E(i,t) + I(i,t) + H(i,t) + R(i,t)\Big) \\
& + \sum_{i\neq j} m_{\text{tr}}(j,i,t)\tau(j,i)S(j,t) - \sum_{i\neq j} m_{\text{tr}}(i,j,t)\tau(i,j)S(i,t), \\
\frac{\mathrm{d}E(i,t)}{\mathrm{d}t} =& \frac{S(i,t)\Big(m_I(i,t)\beta_I(i)I(i,t) + m_H(i,t)\beta_H(i)H(i,t) + m_D(i,t)\beta_D(i)D(i,t)\Big)}{\mathrm{NP}(i,t)} \\
& - \mu_m(i)E(i,t) + \sum_{i\neq j} m_{\text{tr}}(j,i,t)\tau(j,i)\chi_{\epsilon_{\text{fit}}}(E(j,t)) \\
& - \sum_{i\neq j} m_{\text{tr}}(i,j,t)\tau(i,j)\chi_{\epsilon_{\text{fit}}}(E(i,t)) - \gamma_E(i,t)\chi_{\epsilon_{\text{fit}}}(E(i,t)), \\
\frac{\mathrm{d}I(i,t)}{\mathrm{d}t} =& \gamma_E(i,t)\chi_{\epsilon_{\text{fit}}}(E(i,t)) - \Big(\mu_m(i) + \gamma_I(i,t)\Big)I(i,t), \\
\frac{\mathrm{d}H(i,t)}{\mathrm{d}t} =& \gamma_I(i,t)I(i,t) - \Big(\mu_m(i) + (1-\omega(i,t))\,\gamma_{HR}(i,t) + \omega(i,t)\gamma_{HD}(i,t)\Big)H(i,t), \\
\frac{\mathrm{d}R(i,t)}{\mathrm{d}t} =& (1-\omega(i,t))\,\gamma_{HR}(i,t)H(i,t) - \mu_m(i)R(i,t), \\
\frac{\mathrm{d}D(i,t)}{\mathrm{d}t} =& \omega(i,t)\gamma_{HD}(i,t)H(i,t) - \gamma_D(i,t)D(i,t), \\
\frac{\mathrm{d}B(i,t)}{\mathrm{d}t} =& \gamma_D(i,t)D(i,t),
\end{aligned}
\tag{1}
$$

donde $\mu_n(i)$ es la tasa de natalidad (día$^{-1}$) y $\mu_m(i)$ es la tasa de mortalidad (día$^{-1}$) para el país $i$, es decir, el número de nacimientos y muertes, respectivamente, por día y por cápita. Las tasas de contacto efectivo $\beta_I(i), \beta_H(i)$, y $\beta_D(i)$

son parámetros constantes para cada país $i$ representando el número medio de contactos que transmiten la epidemia de una persona en los estados $I$, $H$, y $D$, respectivamente, por día antes de aplicar las medidas de control. El número de personas que pasan de los estados $E$ a $I$, $I$ a $H$, $H$ a $D$, $H$ a $R$ y $D$ a $B$ por día y por cápita se denotan por $\gamma_E(i,t), \gamma_I(i,t), \gamma_{HD}(i,t)$, $\gamma_{HR}(i,t)$, y $\gamma_D(i,t)$, respectivamente, ya que son funciones que dependen tanto del país como del tiempo porque varían con la aplicación de las medidas de control. Para describir la eficiencia de estas medidas de control, se han considerado las siguientes funciones decrecientes:

$$m_I(i,t) = m_H(i,t) = m_D(i,t) = \exp\Big( - \kappa(i) \max(t - \lambda(i), 0.0)\Big).$$

Nótese que estas funciones multiplican las tasas de contacto efectivo de la enfermedad en el Sistema 1, de modo que el número de contactos efectivos se reduce a medida que se mejora la eficiencia de las medidas de control. Esta reducción se gestiona a través del parámetro $\kappa(i) \in [0.0, +\infty)$ (día$^{-1}$), mientras que el parámetro $\lambda(i)$ se refiere al primer día de aplicación de esas medidas de control en el país $i$.

Además, en el Sistema 1, $\omega(i,t)$ es la tasa de fatalidad de la enfermedad, que representa el porcentaje de gente que no sobrevive a la epidemia en cada país $i$ en el instante $t$.

Finalmente, los movimientos migratorios han sido considerados mediante la matriz $(\tau(i,j))_{i,j=1}^{N_{\text{co}}}$, que está compuesta por las tasas de transferencia (día$^{-1}$) de personas desde un país $i$ a otro $j$ expresadas en procentaje (%) de población en el país $i$ por unidad de tiempo. Estas tasas se pueden ver reducidas también por la progresiva aplicación de las medidas de control, por tanto, se multiplican por la función $m_{\text{tr}}(i,j,t)$, que es el producto de la eficiencia de las medidas de control en ambos países $i$ y $j$ involucrados: $m_{\text{tr}}(i,j,t) = m_I(i,t) \cdot m_I(j,t)$.

Como se puede observar en el Sistema 1, para evitar la propagación artifical de la enfermedad debida a valores despreciables de $E(i,t)$, se ha utilizado la siguiente función filtro: $\chi_{\epsilon_{\text{fit}}}(x) = x$ si $x \geq \epsilon_{\text{fit}}$, $\chi_{\epsilon_{\text{fit}}}(x) = 2x - \epsilon_{\text{fit}}$ si $\epsilon_{\text{fit}}/2 \leq x \leq \epsilon_{\text{fit}}$, y 0 en los restantes casos, donde $\epsilon_{\text{fit}} \geq 0$ es un parámetro de tolerancia que toma un valor pequeño.

## 3 Optimización para estimar los parámetros epidemiológicos

La Organización Mundial de la Salud (OMS) proporciona informes periódicos sobre epidemias de interés (tales como el virus del Ébola), en los que refleja el número acumulado de casos y el de muertes en los países afectados. Por lo tanto, asumiendo que se conocen estos datos reales para los $N_{\text{co}}$ países considerados en $N_h$ instantes de tiempo, los denotamos por $\{\text{CC}_{\text{real}}(i,t_j)\}_{j=0}^{j=N_h}$ y $\{\text{CD}_{\text{real}}(i,t_j)\}_{j=0}^{j=N_h}$, respectivamente, donde $i \in \{1,\ldots,N_{\text{co}}\}$ indica el país. Esta información se utiliza, además, para obtener las condiciones iniciales para el modelo epidemiológico. Entonces, cuando utilizamos el modelo para simular la

propagación del brote con una configuración de valores $\phi$ para los parámetros epidemiológicos de interés, éste devuelve una predicción sobre la evolución para el número acumulado de casos $CC^{\phi}(i,t)$ y para el número acumulado de muertes $CD^{\phi}(i,t)$. En particular, para el modelo Be-CoDiS, se calculan como:

$$CC^{\phi}(i,t) = CC(i,0) + \int_0^t \gamma_I(i,t) \cdot I^{\phi}(i,t)\mathrm{d}t, \tag{2}$$

$$CD^{\phi}(i,t) = CD(i,0) + \int_0^t \omega(i,t) \cdot \gamma_{HD}(i,t) \cdot H^{\phi}(i,t)\mathrm{d}t, \tag{3}$$

donde $CC(i,0)$ y $CD(i,0)$ son el número inicial de casos y de muertes disponibles en los informes sobre la enfermedad.

Como el objetivo es que esta evolución se ajuste lo máximo posible a la real, las funciones objetivo se formulan como las diferencias entre los datos de las predicciones del modelo y los reales para cada país $i \in \{1,\ldots,N_{\mathrm{co}}\}$ como se indica a continuación:

$$f_i(\phi) = \frac{\left\|\mathrm{CC}_{\mathrm{real}}(i,t_{\mathrm{f}}) - CC^{\phi}(i,t_{\mathrm{f}})\right\|_{L^2}}{\left\|\mathrm{CC}_{\mathrm{real}}(i,t_{\mathrm{f}})\right\|_{L^2}}, \quad f_{2i}(\phi) = \frac{\left\|\mathrm{CD}_{\mathrm{real}}(i,t_{\mathrm{f}}) - CD^{\phi}(i,t_{\mathrm{f}})\right\|_{L^2}}{\left\|\mathrm{CD}_{\mathrm{real}}(i,t_{\mathrm{f}})\right\|_{L^2}}, \tag{4}$$

donde estos errores relativos se calculan usando la norma $L^2$:

$$\|g(T)\|_{L^2} = \Big(\int_0^T \left(g(t)\right)^2 \mathrm{d}t\Big)^{1/2}.$$

Por lo tanto, el problema multi-objetivo considerado es:

$$\begin{cases} \min\ f_i(\phi), & \forall i \in \{1,\ldots,N_{\mathrm{co}}\} \\ \min\ f_{2i}(\phi), & \forall i \in \{1,\ldots,N_{\mathrm{co}}\} \\ \text{s.t.}\ \ \phi \in \Phi, \end{cases} \tag{5}$$

donde $\phi$ denota el vector compuesto por los parámetros epidemiológicos que se desean estimar y $\Phi$ es el conjunto factible o espacio de búsqueda delimitado por los rangos de esos parámetros.

Para resolver el problema de optimización multi-objetivo, utilizamos el algoritmo llamado *Weighting Achievement Scalarizing Function Genetic Algorithm* (WASF-GA) [2]. Más concretamente, hemos empleado la versión paralela de WASF-GA que propusimos en [3]. Este algoritmo permite obtener un buen conjunto de soluciones de compromiso en una región de interés del espacio objetivo en un tiempo computacional reducido gracias al uso de la computación de altas prestaciones. Al tratarse de un algoritmo genético, explora el espacio de búsqueda mediante los métodos de cruce y mutación encargados de generar nuevos puntos. En particular, en este trabajo, se han considerado el operador de cruce conocido como *Simulated Binary Crossover* (SBX) y la mutación polinómica. Entre la familia de los algoritmos basados en preferencias, a la cual pertenece WASF-GA, éste ha demostrado ser competitivo para abordar problemas con tres o más objetivos.

## 4 Experimentos computacionales y resultados

En este trabajo, la metodología propuesta para ajustar los parámetros epidemiológicos ha sido aplicada al brote del virus Ébola de 2014-2016. En concreto, distintas versiones del Problema (5), variando el número de objetivos y de variables de decisión, han sido resueltas, centrándonos en la predicción de la propagación en aquellos países con el número más elevado de infecciones.

A pesar de que el modelo puede manejar un conjunto de países, los primeros experimentos se han dedicado a validar la metodología de ajuste cuando el objetivo es determinar los parámetros para un sólo país. En estos casos más sencillos, el Problema (5) involucra sólo dos objetivos. Resolviéndolo con WASF-GA, obtenemos un conjunto de puntos de compromiso entre los cuales los epidemiólogos pueden seleccionar el punto que mejor satisfaga sus intereses. En particular, en nuestro análisis, hemos disinguido tres estrategias: (i) seleccionar la solución que proporciona el valor mínimo para el primer objetivo, esto es, el que mejor ajusta el número acumulado de casos; (ii) elegir el que ofrece el mejor compromiso entre los dos objetivos (utilizando la norma Euclídea), y (iii) tomar el punto que proporciona el valor mínimo para el segundo objetivo, esto es, el que mejor ajuste el número acumulado de muertes.

Como WASF-GA es un algoritmo meta-heurístico, hemos realizado varias ejecuciones para asegurar que los resultados no dependen de la aleatoriedad. En particular, hemos repetido 30 veces cada uno de los problemas para un país concreto. Para cada ejecución, hemos registrado los valores de los parámetros correspondientes al punto seleccionado con cada estrategia. Después, hemos calculado los valores promedio y desviación estándar para cada uno de los parámetros. De este estudio, hemos obtenido que la variación de estos valores no es significativa y, por lo tanto, nuestra metodología es robusta. Por ejemplo, el valor promedio de la tasa de contacto efectivo de las personas infectadas en Guinea considerando la estrategia de decisión (ii) es $\beta_I = 0.1907$ y su desviación estándar 4.99E-04. Además, considerando todas las repeticiones de WASF-GA, hemos identificado los valores de los parámetros que proporcionan los peores valores de las funciones objetivo. En la Figura 1, la evolución real del virus del Ébola en Guinea se compara con la evolución predicha por el modelo Be-CoDiS utilizando esos valores para los parámetros. Como puede observarse, incluso considerando la peor solución de WASF-GA, nuestra metodología es capaz de encontrar una configuración para los parámetros tal que el modelo ajusta con precisión el brote real. De hecho, para la estrategia (ii), los errores relativos obtenidos usando (2) y (3) son 1.92E-2 para el número acumulado de casos y 1.70E-2 para el número acumulado de muertes.

En el siguiente estudio realizado, hemos aplicado la metodología de ajuste a un conjunto de 176 países. En este caso, hemos considerado el Problema 5 con las funciones objetivo dadas por (2) y (3) para los tres países donde el virus Ébola ha tenido una mayor virulencia: Guinea, Liberia y Sierra Leona. Además, hemos incluido dos funciones objetivo más para el número acumulado de casos y para el número acumulado de muertes, que consisten en el promedio de sus erro-

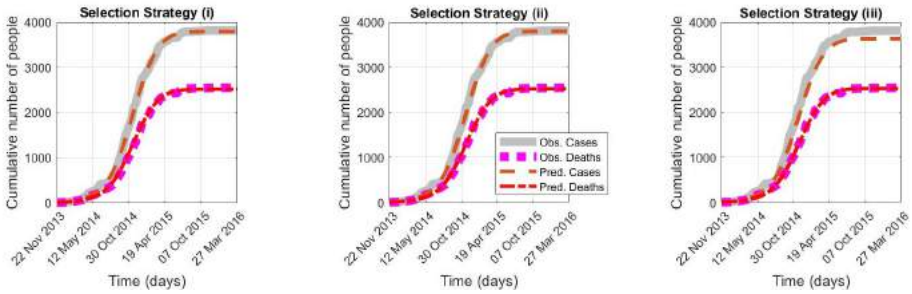



**Fig. 1.** Evolución del número acumulado de casos y de muertes para la epidemia de Ébola en Guinea: datos reales observados (líneas gruesas) y datos obtenidos con el modelo Be-CoDiS (líneas delgadas).

res absolutos para los demás países. Más aún, para evitar fallos en la detección de países infectados, hemos añadido al Problema (5) una función objetivo contando el número de países no-infectados que el modelo predice como infectados (falsos positivos) y otra para el número de países infectados que predice como no-infectados (falsos negativos). Tras analizar los resultados obtenidos como en los anteriores experimentos para un sólo país, podemos concluir que la metodología multi-objetivo propuesta es capaz de encontrar con éxito una configuración de valores para los parámetros del modelo Be-CoDiS que permite describir con precisión la propagación de la enfermedad en todos los países considerados.

## Agradecimientos

El presente trabajo ha sido financiado por el Ministerio de Economía y Competitividad de España mediante los proyectos TIN2015-66680-C2-1-R y MTM2015-64865P; por la Junta de Andalucía, a través del proyecto P12-TIC301, financiado parcialmente por el Fondo Europeo de Desarrollo Regional (FEDER). Este trabajo ha utilizado el servicio de HPC Cirrus del EPCC (https://www.epcc.ed.ac.uk/cirrus) gracias al proyecto HPC-EUROPA3 (INFRAIA-2016-1-730897), financiado por la acción de innovación e investigación europea bajo el programa H2020.

## References

1. Ivorra, B., Ngom, D., Ramos, A.M.: Be-codis: A mathematical model to predict the risk of human diseases spread between countries—validation and application to the 2014–2015 ebola virus disease epidemic. Bulletin of Mathematical Biology **77**(9) (Sep 2015) 1668–1704
2. Ruiz, A.B., Saborido, R., Luque, M.: A preference-based evolutionary algorithm for multiobjective optimization: the weighting achievement scalarizing function genetic algorithm. Journal of Global Optimization **62**(1) (2015) 101–129

3. Ferrández, M.R., Puertas-Martín, S., Redondo, J.L., Ivorra, B., Ramos, A.M., Ortigosa, P.M.: High-performance computing for the optimization of high-pressure thermal treatments in food industry. The Journal of Supercomputing (2018) 1–16

# SECCIÓN II

# Otros trabajos

1. Altamirano Di Luca, Marlon: "Modelo basado en ontología para implementar Web Semántica que apoye la gestión de la información y el conocimiento".
2. Alulema Flores, Darwin Omar: "Una metodología cross-device basada en modelos para IoT".
3. García Salmerón, José Manuel: "Detección de una matriz copositiva mediante la evaluación de las facetas de un simplex unidad".
4. Gómez Navarro, Francisco José: "Avances en el Modelado y Simulación de un Nuevo Concepto de Vehículo Urbano Eléctrico Ligero. Almacenamiento y Distribución de Energía".
5. González Revuelta, Mª Esther: "Interés de los usuarios del Sistema Sanitario en relación al uso de las nuevas tecnologías de la Información y Comunicación (TIC) en la relación médico-paciente y en el seguimiento y evolución de su Proceso Asistencial".
6. Maturana Espinosa, José Carmelo: "Rate Allocation for Motion Compensated JPEG2000".
7. Medina López, Cristóbal: "Atravesando NAT Simétricos en redes P2P mediante predicción de puertos colaborativa".
8. Mena Vicente, Manel: "Una arquitectura de microservicios para componentes digitales en el marco del Internet de las Cosas".
9. Muñoz Rodríguez, Manuel: "Aplicación del IoT en la agricultura intensiva protegida".
10. Ortega López, Luis: "Análisis de imágenes multi-espectrales aplicadas al campo de la agricultura".
11. Sánchez Hernández, José Juan: "Transmisión de secuencias de imágenes JPEG2000 usando actualización condicional y compensación de movimiento controlada por el cliente".
12. Santamaría López, Teresa: "Adaptive Streaming Algorithms and Network Protocols".
13. Wang, Hui: "Improving the performance of vegetable leaf wetness duration models in greenhouses using decision tree learning".

# Modelo basado en ontología para implementar Web Semántica que apoye la gestión de la información y el conocimiento

## *Model based on ontology for semantic web implementer that leans the step of the information and the knowledge*

Marlon Altamirano Di Luca

[1]Master en Seguridad Informática Aplicada. Facultad de Sistemas y Telecomunicaciones. Universidad Estatal Península de Santa Elena, Guayaquil-Ecuador. Mail: marlon.altamiranod@ug.edu.ec

**Resumen:** Los servicios Web se han consolidado como tecnología para el uso de internet, ellos requieren de mecanismos de integración, para establecerse como herramientas tecnológicas que contribuya a la gestión de la información y el conocimiento en disimiles actividades de investigación, desarrollo y propiamente, de las organizaciones. En el presente artículo se propone un modelo basado en ontología, para implementar Web Semántica, que apoye la gestión de la información y el conocimiento almacenada en los repositorios digitales, la cual se administra a través de ontología de dominio, para apoyar la gestión de la información y el conocimiento. En el proceso del modelado de la Web Semántica basada en ontología, se utilizan algoritmos de procesamiento del lenguaje natural, que mejoran la calidad de la información almacenada. Las métricas de precisión y exhaustividad permitieron corroborar la calidad, pertinencia y relevancia de la Web Semántica basada en ontología en la gestión de la información y el conocimiento.

**Palabras clave**: gestión de la información y el conocimiento, ontología, web semántica, repositorios digitales, métricas y algoritmos

**Abstract**: *The web services have consolidated as technology for the use of internet, they require of mechanisms of integration, to establish as technological tools that contributes to the step of the information and the knowledge in dissimilating activities of investigation, develop and properly, of the organizations. At present article proposes a model based on ontology, for web implementer semantic, that leans the step of the information and the stored knowledge in the digital repositories, the who it administers through ontology of dominion, to lean the step of the information and the knowledge. In the process of the modeling of the semantic web based on ontology, use algorisms of prosecution of the natural language, that improve the quality of the stored information. The metrics of precision and exhaustivitad permitted corroborate the quality, pertinence and relevancy of the semantic web based on ontology in the step of the information and the knowledge.*

**Key words**: *Step of the information and the semantic knowledge, ontology, web, digital, metric repositories and algorisms*

## 1. Introducción

Los Servicios Web se han consolidado como una tecnología esencial para la cooperación en Internet, pero requieren mecanismos para su integración, estableciéndose como herramienta tecnológica que contribuya a la globalización y gestión del conocimiento en las actividades organizacionales o en el campo de la investigación, con servicios que mejoren los tiempos de respuesta a los usuarios, en términos de búsquedas eficientes y rápidas.

Esta Web extendida se apoya en lenguajes universales como la lógica descriptiva, los agentes inteligentes y las ontologías, resolviendo las carencias semánticas que hoy hacen difícil y dispendioso el acceso a la información en Internet.

Mientras los estándares relacionados con la composición o colaboración de Servicios Web dan un primer acercamiento, su convergencia con las tecnologías actuales, ofrecen un panorama hacia la obtención de un entorno en el cual la búsqueda, efectividad y ejecución de servicios sea completamente automatizada.

De acuerdo a lo anterior, la Web Semántica es una Web perfeccionada, dotada de mayor significado con el cual, cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla, gracias a una mejor definición de la información.

Con el fin de dar un carácter especializado a los contenidos temáticos, categorizando y catalogando la información a través de la generalización de áreas globales y especializadas, es indispensable hacer uso de las ontologías. El concepto de ontología está basado de tiempo atrás, en la filosofía y recientemente se utiliza en Informática para definir vocabularios que las máquinas puedan entender y que sean especificados con la suficiente precisión como para permitir diferenciar términos y referenciarlos de mejor manera.

Gruber (1993), definió las ontologías como la especificación explícita de una conceptualización, que permite mejorar la recuperación de la información, entre otras actividades derivadas desde la gestión de la información.

Las ontologías generalmente se utilizan para especificar y comunicar el conocimiento del dominio de una manera genérica y son útiles para estructurar y definir el significado de los términos (Fernández, 2015). La utilización de ontologías permite mejorar el proceso de anotación semántica de documentos al dotar los Sistemas de Recuperación de la Información de una base de conocimiento amplia y diversa.

Entre los principales métodos para mejorar la recuperación de la información se describe la anotación semántica de información, que permite transformar el texto original obtenido al rastrear la web en un documento enriquecido a partir de incluir diversos términos que mejoran la comprensión de la información almacenada (Rodríguez, 2014; Legaz, 2015; Otero, 2017). Las anotaciones semánticas reducen el espacio entre el lenguaje natural y la representación computacional de la información enlazando los términos de un documento con su representación semántica en la ontología que representa el conocimiento de forma estructurada (Otero, 2017). En un documento un término puede tener varios significados o varios términos pueden referirse a un mismo concepto añadiendo complejidad el procesamiento de la información.

Según Otero (2017) para disminuir la ambigüedad de la información y aumentar la precisión de los resultados de búsquedas, los sistemas de anotación semántica deben explotar el contexto del término analizando su significado en el documento. La utilización de ontologías en el proceso de anotación semántica permite identificar mejor el significado de cada término del documento.

Las ontologías incluyen definiciones de conceptos básicos relacionados con un dominio, así como las relaciones entre ellos, de tal forma que los computadores pueden codificar el conocimiento y también el conocimiento extendido, haciendo reutilizable el conocimiento. Integrando todos los aspectos anteriores, las organizaciones y entidades, cuentan con herramientas tecnológicas eficientes de comunicación y organización de información, de

tal forma que el conocimiento organizacional podrá ser gestionado, categorizado, retroalimentado y ampliamente divulgado.

Para realizar la implementación de ontologías en los servicios web, se requiere hacer uso del lenguaje OWL (Ontology Web Language), el cual proporciona terminologías interpretables por la Web para la creación de estructuras ontológicas, brindando integración e interoperabilidad de datos descriptivos para el trabajo entre diversas comunidades (Guarino, 1995).

OWL proporciona las siguientes capacidades a las ontologías:

- Capacidad de ser distribuidas a través de varios sistemas.
- Es escalable a las necesidades de la Web.
- Es compatible con los estándares Web de accesibilidad e internacionalización.
- Es abierto y extensible.
- Da utilidad de las Ontologías para la Web.

Especialistas de la Facultad de Sistemas de la Universidad Estatal Península de Santa, identificaron los principales casos de uso de ontologías en la Web; se realizó un estudio sobre los servicios implementados con lenguajes de ontologías para Web poco avanzados, de donde se obtuvo la siguiente clasificación:

- Portales Web
- Repositorios Digitales
- Reglas de categorización utilizadas para mejorar la búsqueda.
- Colecciones multimedia.
- Búsquedas basadas en contenido para medios no textuales.
- Administración de Sitios Web Corporativos.
- Organización taxonómica automatizada de datos y documentos.
- Asignación entre Sectores Corporativos.
- Documentación de Diseño.

- Explicación de partes "derivadas" (Ej.: el tornillo de una pieza mecánica).
- Administración explícita de restricciones.
- Agentes Inteligentes.
- Expresión de las preferencias y/o intereses de los usuarios.
- Mapeo de contenidos entre sitios Web.
- Composición y descubrimiento de Servicios Web.
- Administración de derechos y control de acceso.

Por otra parte, se detecta que la información se encuentra en diferentes formatos expresada en lenguaje natural, lo que para mejorar su procesamiento se hace necesario la realización de anotaciones semánticas a los documentos indexados (Rodríguez, 2014; Blandón, 2017). La anotación de información es el proceso que permite asociar conceptos, relaciones, comentarios o descripciones a un documento o fragmento de texto para mejorar el proceso de inferencia de conocimiento (Oliveira y Rocha, 2013; Rodríguez, 2014; Vállez, 2015).

La identificación de los elementos antes referido son favorables para el desarrollo de un modelo basado en ontología para implementar Web Semántica que apoye los procesos de gestión de la información en la universidad Estatal Península de Santa Elena, en Guayaquil Ecuador.

## 2. Materiales y métodos

Los materiales y métodos utilizados en el desarrollo de un modelo basado en ontología, para implementar Web Semántica que apoye los procesos de gestión de la información en la universidad Estatal Península de Santa Elena, en Guayaquil Ecuador fueron:

- **Analítico-sintético**: A partir del análisis de los referentes teóricos y la bibliografía relacionada con la investigación se descompuso el problema científico en elementos por separado para profundizar su estudio y sintetizarlos en la propuesta de solución.
- **Hipotético-deductivo**: Mediante la observación y el análisis del fenómeno en cuestión y a través de reglas lógicas de deducción, se formuló una hipótesis que se comprouba en el proceso de validación del modelo propuesto.
- **Histórico-lógico**: Para determinar los antecedentes, tendencias y particularidades de la recuperación de información con anotación semántica, en función de comprender mejor el objeto de estudio de la investigación.
- **Análisis documental**: Se realizaron consultas de libros, artículos científicos, proyectos de investigación para el estudio de los referentes teóricos.

Por otra parte, se tuvo en cuenta aspectos derivados del análisis documental, así como la importancia de la utilización de repositorios digitales, para gestionar y compartir información, en aras de implementar una Web

Semántica, que contribuya a que la información sea más viable, dado que la misma estará mejor organizada y estructurada.

El grado de utilidad de los procesos de revisiones en la gestión de la información, que se procesa y gestiona en los repositorios digitales, las funcionalidades del repositorio digital para garantizar la calidad de los recursos que se publican, el estado de utilización de la información que se gestiona en el repositorio digital y de los estándares de catalogación, empaquetamiento e interoperabilidad en las universidades ecuatorianas de Guayaquil.

Basado en lo antes referido se propone la concepción metodológica que guía el proceso de desarrollo del modelo, teniendo en cuenta; la estructura, cualidades, premisas y principios del modelo a desarrollar. Dicha concepción metodológica se presenta en la figura 1.

El modelo propuesto, tiene una estructura abierta debido a que establece intercambio con el entorno general y específico. En el caso del entorno específico permite el intercambio con los profesores de las universidades ecuatorianas en Guayaquil y con los especialistas en Tele formación de la Universidad objeto de estudio. La interacción con el entorno general se establece de forma indirecta a través de factores tecnológicos y políticos legales.

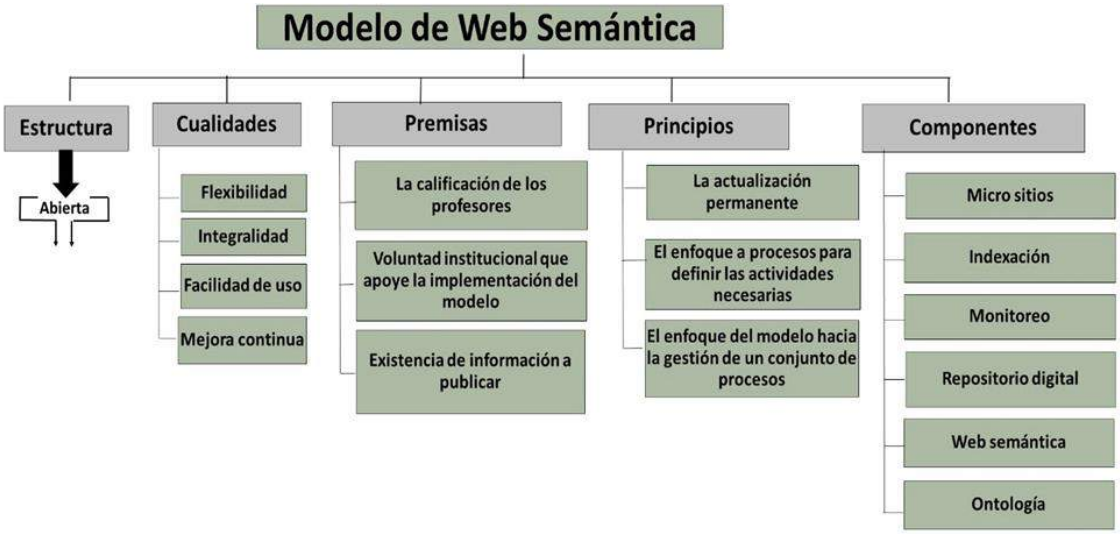



**Figura 1.** Estructura, componentes, premisas, cualidades y principios del modelo para implementar Web Semántica sobre repositorios digitales. **Fuente:** Elaboración propia.

Las cualidades que distinguen el modelo son:

- ✓ Flexibilidad
- ✓ Integralidad
- ✓ Facilidad de uso
- ✓ Mejora continua

Las premisas con vistas a la aplicación del modelo propuesto son:

1. La **existencia** información a publicar útil para catalogar la misma y conocer la mejor forma de indexarla para localizarla a través de los motores de búsqueda.

2. La **calificación** de los profesores necesaria para el uso eficiente del modelo propuesto para trabajar con la información que se publica en los repositorios digitales.

3. La **voluntad** institucional para apoyar la aplicación del modelo.

**El modelo se sustenta bajo los principios de:**

1. La actualización permanente mediante la retroalimentación de la información que nutre al modelo.
2. El enfoque a procesos para definir las actividades necesarias que permitan alcanzar el resultado deseado, identificar las posibles entradas y salidas, así como la evaluación de la información a publicar.
3. El enfoque hacia la gestión de un conjunto de procesos para identificar, entender y gestionar la información necesaria a consultar a través de los repositorios digitales.

De acuerdo con la concepción metodológica para el desarrollo del modelo, basado en ontología útil para implementar Web Semántica, que apoye la gestión de la información y el conocimiento, se establecen criterios para su implementación, con el propósito de obtener resultados favorables en las búsquedas de información que se realizan en los repositorios digitales de la Universidad Estatal península de Santa Elena en Guayaquil – Ecuador.

## 3. Resultados

Sobre los principios y cualidades establecidos se desarrolló el modelo basado en ontología para implementar una Web Semántica en el repositorio digital de la Universidad Estatal Península de Santa Elena Guayaquil – Ecuador, figura 2. Su objetivo es brindar a los profesores y especialistas de tele formación, de las universidades ecuatorianas y en particular las de Guayaquil, información ordenada que sea fácil de encontrar a través del uso de cualquier motor de búsqueda, información previamente gestionada, útil porque facilita a través de la información, obtener conocimiento, socializarlo y a su vez reutilizarlo en las diferentes acciones que se realizan a nivel de instituciones educacionales.

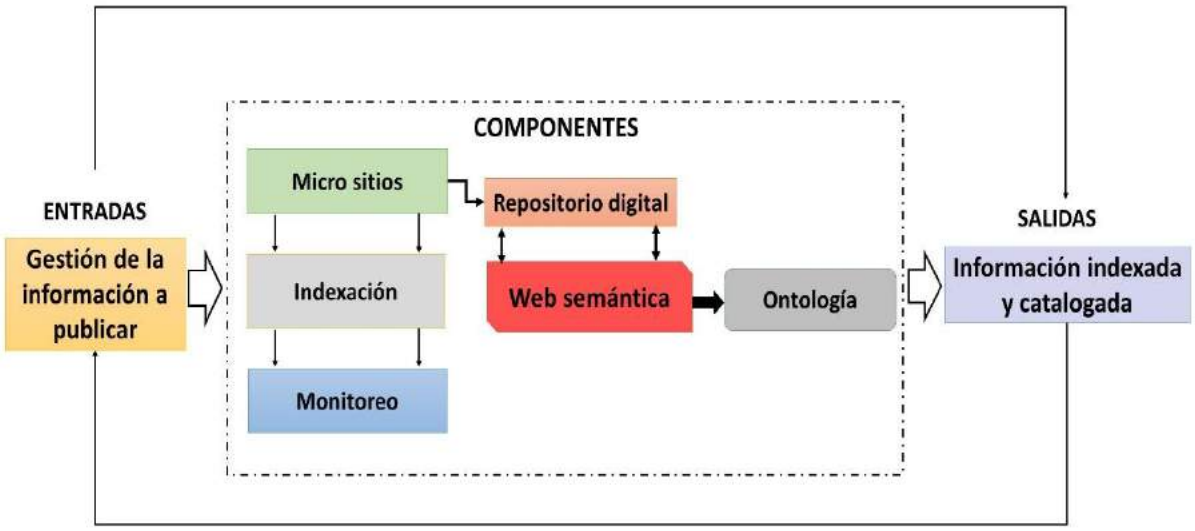



**Figura 2**. Modelo basado en ontología para implementar una Web Semántica en el repositorio digital de la Universidad Estatal Península de Santa Elena Guayaquil – Ecuador. **Fuente**: Elaboración propia.

Las entradas del modelo lo constituyen la información que se gestiona de los departamentos de la Universidad Estatal Península de Santa Elena, Guayaquil Ecuador. Como salida se obtiene información indexada y catalogada del repositorio digital que lo soporta una Web Semántica basada en ontología de dominio, útil para el apoyo a la gestión de la información y el conocimiento almacenada en los repositorios digitales.

El funcionamiento de los componentes del modelo, se basa en el supuesto que posee esta nueva Web, en particular la capacidad de construir una base de conocimiento sobre las preferencias de los usuarios y que, a través de una combinación entre capacidad de comprensión e información disponible en Internet, que sea capaz de atender de forma exacta las demandas de información por parte de los usuarios en relación.

En este modelo la Web Semántica aporta un camino para razonar en la Web al ser una infraestructura basada en metadatos (datos altamente estructurados que describen información), extendiendo así sus capacidades.
No se trata de una inteligencia artificial que permita a los servidores Web entender las palabras de los usuarios, es sólo la construcción de una habilidad dispuesta en una máquina, con el fi n de resolver problemas bien definidos, a través de operaciones igualmente bien definidas que se llevarán a cabo sobre datos existentes.

Para esta labor, la Web Semántica utiliza *Resource Description Framework* (RDF) y *Ontology Web Language* (OWL), los cuales son dos estándares que ayudan a convertir la Web en una infraestructura global en la que es posible compartir, y reutilizar datos y documentos entre diferentes tipos de usuarios (Cabral, Domingue, Motta, 2004).

RDF proporciona información descriptiva simple sobre los recursos que se encuentran en la Web y que se utiliza, por ejemplo, en catálogos de libros, directorios, colecciones personales de música, fotos, eventos, etc.
OWL es un mecanismo para desarrollar temas o vocabularios específicos que se encuentran asociados a diversos recursos; lo que hace OWL es proporcionar un lenguaje para definir ontologías estructuradas que pueden ser utilizadas a través de diferentes sistemas.

La ontología, se encargan de definir los términos utilizados para describir y representar un área de conocimiento, son aplicadas por los usuarios, las bases de datos y las herramientas que necesitan compartir información específica, (Colomb, 2002). La ontología incluye definiciones de conceptos básicos en un campo determinado y establece la relación entre ellos.

Los archivos RSS contienen metadatos sobre fuentes de información que han sido especificadas como de interés por parte de los usuarios, su función principal es notificar cambios acerca de los recursos seleccionados, sin necesidad de comprobar directamente ingresando a la página Web, es decir, los RSS notifican de forma automática cualquier cambio que se realice en esos recursos de interés.

Dentro de FOAF se puede destacar FOAF-a-Matic, que se trata de una aplicación Javascript que permite crear una descripción FOAF de uno mismo; con esta descripción, los datos personales serán compartidos en la Web pasando a formar parte de un motor de búsqueda donde será posible descubrir información acerca de una persona en concreto y de las comunidades de las que es miembro de forma sencilla y rápida.

Los buscadores semánticos son un ejemplo más de aplicaciones específicas, cuyo objetivo es satisfacer las expectativas de búsqueda de usuarios que requieren respuestas precisas.

Descritos los elementos fundamentales del modelo, es de destacar que, en cuanto a los términos de búsqueda, la web semántica no genera un impacto visual que lo diferencie de otras técnicas preexistentes, como son los buscadores convencionales. Sin embargo, su real potencial se encuentra en la dinámica de su funcionamiento, su concepto, sus fundamentos, convirtiéndose en un mecanismo eficiente de clasificación, organización y dinamizador de la información.

## Conclusiones

El modelo propuesto para implementar Web Semántica que apoye la gestión de la información y el conocimiento permite mejorar la relevancia de los resultados de búsqueda brindados a los usuarios en un Sistema de Recuperación de información, lo que contribuye a la toma de decisiones individuales y de las organizaciones.

Para mejorar la precisión y relevancia de los resultados de búsqueda brindados a los usuarios se realizan anotaciones semánticas en los documentos, lo que exige una preparación de los especialistas vinculados a la gestión de información en las universidades y centros de investigación.

Las métricas de Precisión y Exhaustividad de los componentes incorporado al modelo, triangulados con los resultados de la consulta a expertos, demostraron que los resultados obtenidos fueron satisfactorios

## Referencias

1. BLANDÓN, J. C. (2017). Extracción de instancias de una clase desde textos en lenguaje natural independientes del dominio de aplicación. Tesis Doctoral, Universidad Nacional de Colombia - Sede Medellín, Colombia.
2. CABRAL Liliana, DOMINGUE John, MOTTA Enrico, (2004). Approaches to Semantic Web Services: An Overview and Comparisons. IAM, University of Southampton, Southampton, UK.

3. COLOMB Robert (2002). The physical being of institutional facts - National Research Council - Institute of Biomedical Engineering - ISIB-CNR - corso Stati Uniti, 35127 Padova.
4. FERNANDEZ, J.A (2015) Ontología, funciones y discurso en el videojuego. Revistas académicas. Humanidades. Universidad de Costa Rica. Vol. 7 no 1., doi.org/10.15517/h. v7i1.27641
5. GRUBER, T. R. (1993). A translation approach to portable ontology specifications, Knowledge Acquisition, vol. 5 no. 2, pp. 199–220.
6. GUARINO Nicola, (1995). Formal Ontology, Conceptual Analysis and Knowledge Representation - LADSEB-CNR, National Research Council, Padova, Italy.
7. LEGAZ, M. D. (2015). Integración de información biomédica basada en tecnologías semánticas avanzadas, Tesis doctoral, Universidad de Murcia.
8. OLIVEIRA, P., ROCHA, J. (2013). Semantic annotation tools survey. En Computational Intelligence and Data Mining (CIDM), IEEE Symposium on. IEEE, pp. 301-307. DOI: 10.1109/CIDM.2013.6597251
9. OTERO, E. N. (2017) Descubrimiento de grafos en datos enlazados para la anotación semántica de documentos. Tesis doctoral, Universidad de Santiago de Compostela, Galicia, España.
10. RODRÍGUEZ, M. Á. (2014) Extracción semántica de información basada en evolución de ontologías. Tesis doctoral, Universidad de Murcia, España, 2014.
11. RODRÍGUEZ, M. A. et al. (2014). Ontology-based annotation and retrieval of services in the Cloud. Knowledge-Based Systems, 2014. 56, pp. 15-25.
12. RODRÍGUEZ, M.A. et al. (2014). Creating a Semantically-Enhanced Cloud Services Environment through Ontology Evolution. Future Generations in Computer Systems, 32, pp. 295–306.
13. VÁLLEZ, M. (2015). Exploración de procedimientos semiautomáticos para el proceso de indexación en el entorno web, Tesis doctoral, Universidad de Barcelona, España.

# Una metodología basada en modelos para conectar dispositivos heterogéneos del Internet de las Cosas y de la Televisón Digital

Darwin Alulema[1,2]

[1] Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador
[2] Applied Computing Group, University of Almería, Spain
doalulema@sespe.edu.ec

**Resumen** . El desarrollo de Internet ha provocado una nueva revolución, que modifica la forma en que vivimos, trabajamos y nos relacionamos, las personas y los objetos, lo que ha dado lugar a la era del Internet de las Cosas (IoT). Por lo cual el número de aplicaciones ha incrementado y cubre una amplia variedad de escenarios como *Smart City*, *Smart Agro*, *Smart Building*, *Smart Home* o *Smart Health*, entre otros. Sin embargo, esta diversidad tiene la dificultad de tener que coordinar la interacción entre dispositivos y plataformas heterogéneas. Por consiguiente se requiere que los desarrolladores deban tener un alto grado de conocimiento de cada una de las tecnologías empleadas. Por esta razón, proponemos una metodología basada en modelos para facilitar el proceso de desarrollo. La propuesta permite modelar una arquitectura que integra plataformas heterogéneas para IoT, que integra la TV digital (DTV), Smart Phones y otros nodos hardware, como por ejemplo, Google Home y Alexa o los creados por medio de placas de desarrollo tipo Arduino o Raspberry. Para este objetivo, se ha diseñado un conjunto de herramientas gráficas que permiten modelar y generar código de forma automática por medio de transformaciones modelo-a-texto (M2T). Para demostrar el funcionamiento de la propuesta se han desarrollado escenarios de prueba en el ámbito del Smart Health y Smart Home.

**Keywords:** Model-Driven Engineering (MDE) · Domain Specific Language (DSL) · Internet of Things (IoT) · Digital Television (DTV) · Smart Home · T-Health

## 1. Introducción.

El Internet de las Cosas (IoT), es una red global de objetos que interactúan entre ellos y las personas [13]. El cocimiento que ofrecen los objetos al interconectarse permiten establecer nuevas aplicaciones y escenarios [10]. De hecho, se espera que miles de millones de dispositivos desempeñen un papel importante en la red futura, trayendo datos del mundo físico al mundo de los contenidos y servicios digitales [3]. Además estos objetos al incorporar sensores permiten recopilar continuamente datos de su entorno y de las personas, sin interferir con

las actividades diarias [16]. Sin embargo, este crecimiento presenta el problema de la diversidad de las plataformas, la eficiencia de código, interacción con recursos del dispositivo [7], o corto tiempo de comercialización. Esto hace que los desarrolladores se enfrenten ante un gran desafío al diseñar aplicaciones que se ejecuten en diferentes plataformas [5].

Al considerar la interoperabilidad, existen dos problemas principales [8]: la complejidad del desarrollo de software y la fragmentación de plataformas. Como una solución a este problema, los conceptos de "Cross-device" [12] y la arquitectura dirigida por modelos (MDA) [5] que se pueden aplicar para acelerar el proceso de desarrollo de aplicaciones [15].

Para acelerar el diseño de aplicaciones de IoT interoperables, capaces de comunicarse entre sí sin dificultad, proponemos una metodología basada en modelos para la abstracción de aplicaciones para IoT. Las principales aportaciones son: un metamodelo que permite abstraer un escenario para IoT, un editor gráfico para facilitar el proceso de construcción de aplicaciones de IoT que permite el diseño de la arquitectura del sistema, un editor gráfico para la construcción de aplicaciones de DTV, un editor gráfico para aplicaciones de nodos de Hardware. Además, se ha realizado un prototipo para la generación semiautomática de código para el estándar ISDB-Tb para DTV y para plataformas de desarrollo Arduino.

El resto de este artículo se organiza como sigue. La sección 2 revisa algunos trabajos relacionados. La sección 3 presenta un lenguaje específico de dominio (DSL), un editor gráfico y una transformación de modelo a texto (M2T) para la generación de códigos. La sección 4 muestra la viabilidad del enfoque a través de un escenario de prueba. Finalmente, el trabajo futuro y las conclusiones se extraen en la Sección 5.

## 2. Trabajos relacionados.

La principal herramienta del paradigma basado en la evidencia, es la revisión sistemática de la literatura que proporciona un marco para la búsqueda de la literatura relacionada, permitiendo categorizar, clasificar y realizar análisis temáticos [4]. Esta técnica se ha aplicado en el proceso de búsqueda de trabajos relacionados con el IoT. Para el proceso de Revisión Sistemática hemos propuesto las siguientes preguntas:

- RQ1: ¿Cuáles son las plataformas utilizadas para IoT?
- RQ2: ¿Qué mecanismos se utilizan para el diseño de arquitecturas?
- RQ3: ¿Qué tecnología existe para la interconexión entre plataformas?
- RQ4: ¿Qué tecnología existe para el modelado de software?

El motor de búsqueda seleccionado fue la base de datos Scopus. Este indexa varios catálogos de publicaciones IEEE Xplore, Science Direct, Springer Link, entre otros. Durante la búsqueda, todos los artículos publicados entre 2004 y 2018 se han tenido en cuenta. Los estudios incluidos deben cumplir las siguientes condiciones: a) Completar los artículos, b) Que pertenezcan a la rama Informática, y c) Escrito en inglés.

Existen diferentes tipos de escenarios en los que el IoT está presente. Uno de estos son los servicios de atención médica, como el propuesto en [16], en el que los autores sugieren un sistema para monitorear la actividad física en tiempo real en una casa inteligente. Esta solución utiliza el aprendizaje multitarea, los diccionarios y el aprendizaje de razonamiento basados en reglas para observar y cuantificar los cambios en las lecturas de los sensores instalados en entornos domésticos, con el fin de llevar a cabo un monitoreo continuo del comportamiento de los residentes y detectar cualquier evento anormal para la asistencia médica temprana. La principal diferencia con nuestro enfoque, es el uso de Model-Driven Engineering (MDE) para el diseño de aplicaciones. Nuestro enfoque, mediante el uso de modelos genéricos, se puede utilizar en diferentes dominios. De esta manera, MDE permite a los desarrolladores resolver problemas específicos de integración, considerando y estudiando una solución general independiente de la tecnología.

La diversiaad de aplicaciones que tiene el IoT ha promovido el desarrollo de muchas plataformas para IoT, algunas de las más populares son: Amazon AWS, ARM Bed, Microsoft Azure IoT, Google Brightness / Weave, Calvin Ericsson, Apple HomeKit, Eclipse Kura y Samsung SmartThings, que se utilizan para el desarrollo de aplicaciones inteligentes [1]. Sin embargo hay propuestas no comerciales como la que desarrollan en [7], en la cual sugieren un hogar inteligente que utiliza sistemas ciberfísicos para monitorear y medir las actividades físicas. En este caso, los autores proponen una arquitectura basada en componentes para incorporar sistemas heterogéneos ciberfísicos. Cada dispositivo se puede administrar como un componente encapsulado dentro del concepto de IoT. El enfoque permite la interoperabilidad mediante la aplicación de una representación de componentes homogénea que proporciona funciones de comunicación a través de sockets web y la implementación de pasarelas. Los autores en [14] centran su trabajo en sistemas domésticos inteligentes, a través de interfaces estandarizadas. La perspectiva basada en modelos en el dominio de los Sistemas de automatización de edificios (BAS), permite generar varios tipos de artefactos de texto para el estándar OBIX, con acceso a tecnologías de comunicación como BACnet, KNX, EnOcean o M-Bus. En contraste, nuestra propuesta modela a un nivel más alto de abstracción, lo que permite que se utilicen en otras aplicaciones además de la domótica. El uso de MDE en un entorno heterogéneo, como el IoT, permite acelerar el proceso de desarrollo porque estandariza los parámetros de los componentes cibernéticos. De esta forma, otros desarrolladores pueden crear nuevas aplicaciones. Además, nuestro enfoque utiliza transformaciones M2T, lo que hace posible automatizar el proceso de diseño de software para los componentes de hardware.

Además, debido a los diferentes escenarios de aplicación para IoT, las tecnologías subyacentes también han evolucionado. Una tecnología que se ha beneficiado de este desarrollo son las redes de sensores inalámbricos (WSN). Por esta razón, los autores de [11] proponen una plataforma de modelado para una arquitectura de desarrollo y análisis de redes WSN por generación de código. La plataforma propuesta consta de tres lenguajes de modelado para describir

vistas de arquitectura específicas de una red de sensores WSN: (a) lenguaje de modelado de arquitectura de software, (b) lenguaje de modelado de nodos, y (c) lenguaje de modelado del entorno. La principal diferencia con nuestro enfoque es el uso del protocolo 802.11 (WiFi), que permite que el control y la supervisión de los nodos de hardware se realicen de forma remota, ya que los nodos se tratan como otro elemento de los servicios. Otra tecnología impulsada por el IoT son los sistemas operativos de plataformas de hardware. En este contexto, los autores de [9] proponen una herramienta basada en modelos para desarrollar y configurar aplicaciones para Contiki OS, pero específicamente para la pila de protocolos de red. A diferencia de este proyecto, nuestro enfoque utiliza controladores que no ejecutan sistemas operativos e incluye sensores y actuadores.

En [13] los autores realizaron un estudio para determinar las líneas de investigación del modelado y la generación automática de código para aplicaciones de redes de sensores inalámbricos. Sus resultados muestran la prevalencia de las propuestas para generar código basado en enfoques MDE, destacando sus beneficios y características adecuadas para la generación de código. También mencionan la relevancia de los teléfonos inteligentes para las nuevas aplicaciones de IoT, ya que incluyen varios protocolos de comunicación y tienen una gran cantidad de sensores. Para el caso, como se propone en [6], el autor presenta una arquitectura independiente del dispositivo que separa las aplicaciones de los dispositivos y permite el desarrollo de la aplicación. Para demostrar el enfoque, el autor introduce un escenario para implementar la arquitectura orientada a servicios para que los dispositivos móviles accedan a los recursos. Además, el crowdsourcing se utiliza para determinar el rendimiento de la aplicación en relación con el retraso para acceder a los servicios. La principal diferencia con nuestro enfoque es el uso de MDE para desarrollar modelos de nodos de hardware que se interconectan con otras aplicaciones utilizando una arquitectura de servicios. Por el contrario en [8], los autores proponen una herramienta gráfica basada en modelos para sistemas IoT, en la cual los sensores (acelerómetros, GPS, presión, luz, temperatura, gravedad o proximidad) de un Smart Phone pueden aportar información a un sistema IoT. Aún cuando nuestra propuesta emplea técnicas de modelado igual que en los trabajos mencionados, se diferencia en que incorpora a la DTV.

## 3. Metodología propuesta.

Esta sección describe la metodología propuesta para el diseño de aplicaciones Cross-device de acuerdo con MDE. La propuesta requiere seis procesos divididos en dos etapas: una para la especificación y otra para el desarrollo. En estos procesos se identifican dos actores, que interactúan con el sistema en diferentes etapas:

- El Ingeniero: es el individuo que tiene conocimiento técnico y experiencia en el dominio específico, responsable de traducir las características del sistema;
- El desarrollador: la persona que configura y genera el código específico de la aplicación.

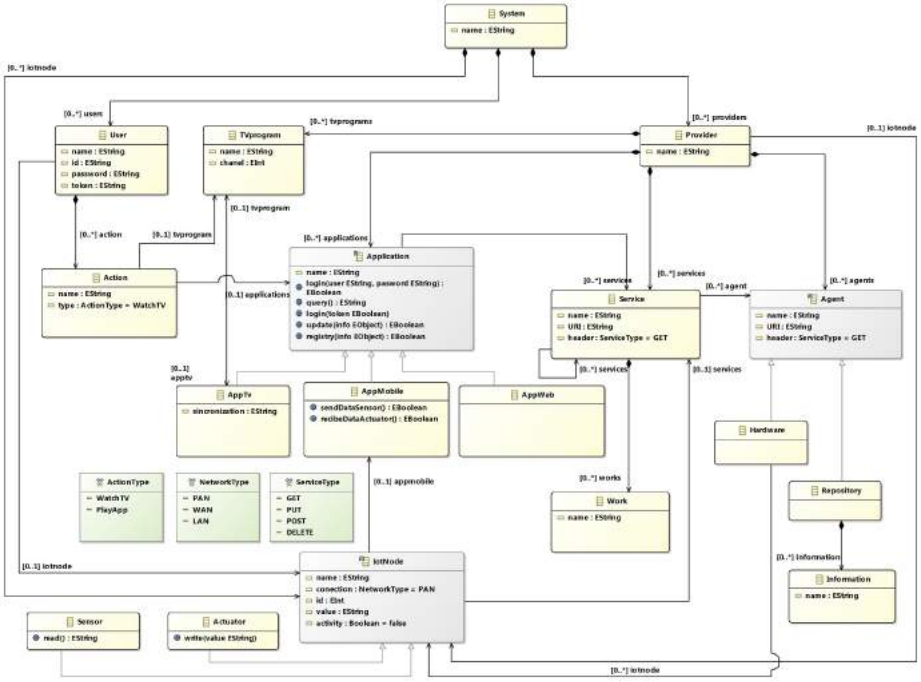



**Figura 1.** Metamodelo de la Arquitectura

La etapa de especificación consta de tres procesos, en los cuales se desarrolla el DSL y se implementa un proceso de transformación M2T. En el lado del rol del ingeniero, el genera el metamodelo, que define la sintaxis abstracta del lenguaje de acuerdo con las características de las aplicaciones. Luego puede realizar: a) Transformación M2T, para generar el código fuente; o b) Desarrollar el editor gráfico, que corresponde a la representación gráfica del DSL.

La etapa de desarrollo consta de tres procesos consecutivos (lado del desarrollador), en los cuales se realiza la aplicación y se genera el código fuente. El primer proceso corresponde a la función establecida en el paso anterior, donde un modelo se construye utilizando el editor gráfico definido en la etapa de especificación. Este editor gráfico se utiliza para describir el escenario de la aplicación y generar el modelo específico del escenario. Después, se realiza la transformación M2T mediante el modelo generado. Este proceso crea automáticamente el código fuente de la DTV y el NodoIoT.

La Figura 1 describe el Metamodelo propuesto para la arquitectura Cross-Device En este caso se diferencian: a) Usuario, es quien consumirá los servicios ofrecidos por el proveedor. La interacción que tiene con el sistema depende de su capacidad de conectividad, b) Proveedor, es el que ofrece el servicio. La interacción que tiene con el sistema depende de su capacidad para desarrollar, difundir y ofrecer un servicio, y c) NodoIoT, es un actor no humano que se puede encontrar en el lado del usuario cuando consume un servicio final y contribuye a su información, para tener una experiencia más enriquecida. Cuando el nodo está del lado del proveedor, sirve para ofrecer un servicio que puede ser consumido por el usuario. Puede cumplir la función de un sensor o actuador, o ambos al mismo tiempo.

La Figura 2 describe el Metamodelo propuesto para las aplicaciones de DTV. Para definir cuales son las más comunes, se ha considerado los ejemplos propuestos en [2]: a) Banner, ordena verticalmente los botones en el lado izquierdo de la pantalla y muestra el texto consultado en la parte inferior de la pantalla. Esta interfaz se observa principalmente en las redes de noticias para mostrar mensajes cortos; b) Accordion, ordena verticalmente los botones en el lado derecho de la pantalla y muestra el contenido invocado en un campo. Esta interfaz se ve principalmente en las aplicaciones de la tienda para expandir información, y c) Frame, utiliza el 75 % de la pantalla al reducir la señal de video del televisor en un campo pequeño y ordena en la información de área disponible de la aplicación y todos los recursos llamados por los botones . Esta interfaz se ve principalmente en aplicaciones que requieren más información como en los programas de cocina.

Los elementos visuales a los que acceden las aplicaciones son: a) Video, una señal de TV que impregna la aplicación que se transmite, y videos cortos, que son recursos multimedia llamados por los botones; b) Text, existe la posibilidad de obtener información localmente (cuando está dentro del mismo carrusel de datos), o remota (cuando se consulta desde un servicio web); c) Button, los elementos componentes tradicionales de la DTV (rojo, verde, amarillo y azul) para la interacción con los usuarios; d) Imagen, los archivos de imagen que se muestran cuando son llamados por los botones; y e) Servicio, que representa el

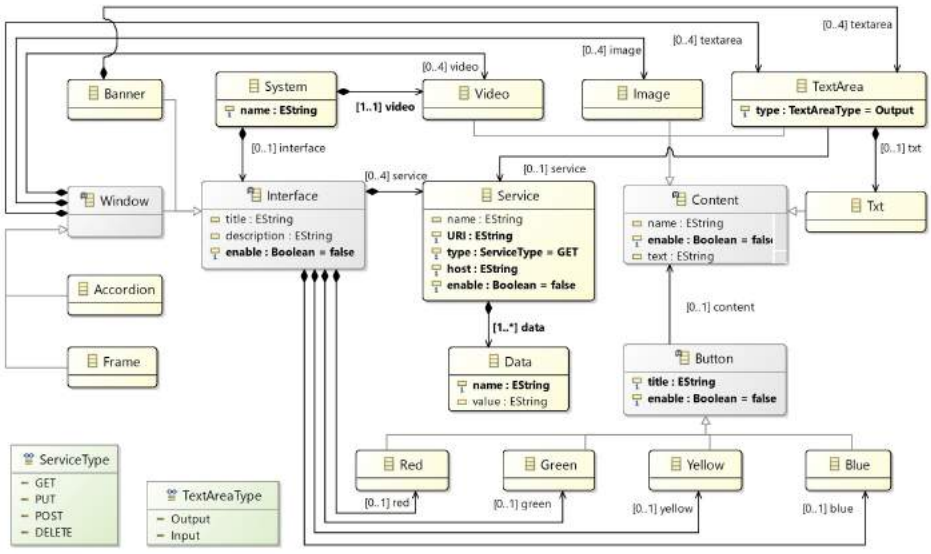



**Figura 2.** Metamodelo IoTV.

mecanismo para el consumo de servicios web y el almacenamiento temporal de esta información en la televisión.

La Figura 3 describe el metamodelo que describe los servicios y los nodos de hardware construidos en una plataforma de desarrollo (*p. Ej., Arduino y Raspberry Pi*). Estos nodos también tienen conectividad (*p. Ej.*, USB, serial, Bluetooth, WiFi y Ethernet) para conectarse con otros dispositivos o servicios cercanos. La plataforma también permite controlar múltiples tipos de componentes analógicos o digitales (*i.e.*, sensores y actuadores). Sin embargo, el metamodelo propuesto se encuentra en un nivel más alto de abstracción, por lo que no se consideran muchos detalles de las interfaces de los puertos y componentes. En este nivel, los puertos se han representado como entradas y salidas, y los componentes como sensores y actuadores. Esta representación permitió simplificar el proceso de conexión de los componentes y el controlador.

## 4. Escenario de estudio de caso.

Para el desarrollo de aplicaciones Cross-Device, se han creado tres editores gráficos basados en los metamodelos propuestos anteriormente. La Figura 4 muestra una captura de pantalla de los editores, con los que se han analizado tres escenarios distintos para verificar su funcionamiento. En la Figura 4.a se observa el escenario de un Parking inteligente en el cual se han determinado etapas de control y potencia. En la Figura 4.b se ha desarrollado un escenario en

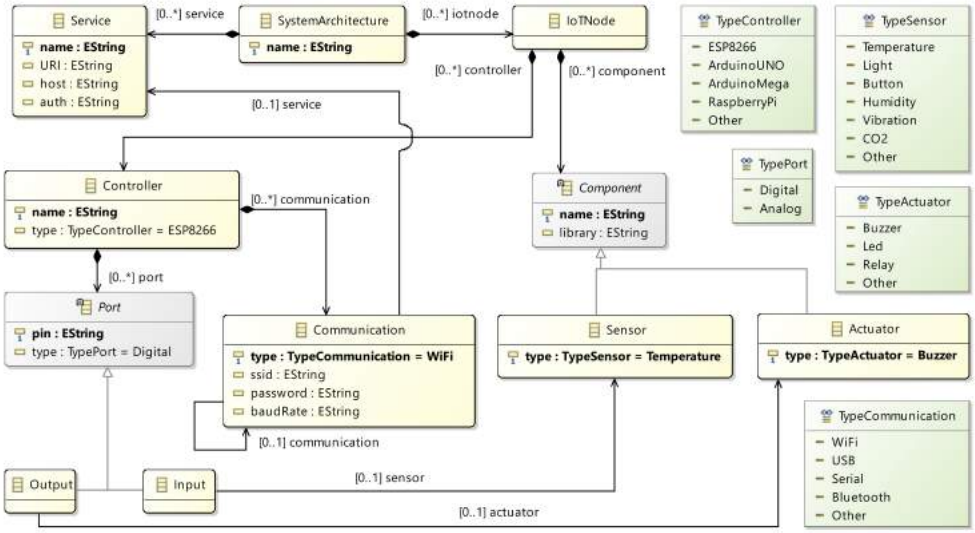



**Figura 3.** Metamodelo NodoIoT.

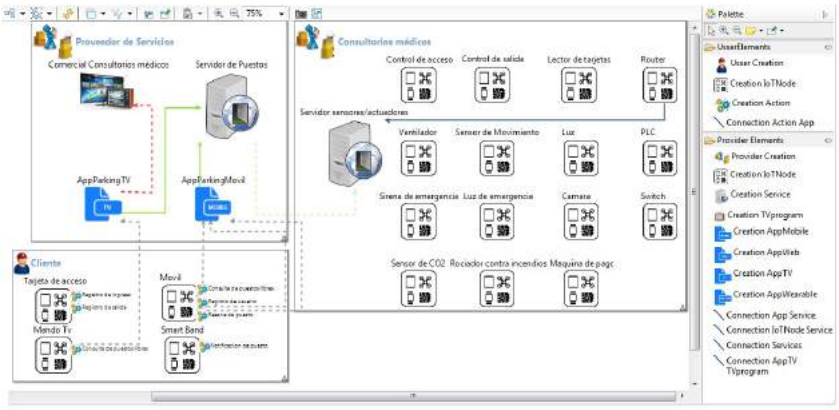
el cual el usuario accede a la información de una estación meteorológica desde una aplicación de DTV.

Además, la herramienta es capaz de generar de forma automática el código de la aplicación en el estándar ISDB-Tb de televisión. Por último en la Figura 4.c se ha diseñado una Smart Home, al integrar varios sensores y actuadores que interactúan con las personas. En este caso la herramienta es capaz de generar el código para tarjetas Arduino para que puedan acceder a los servicios web.

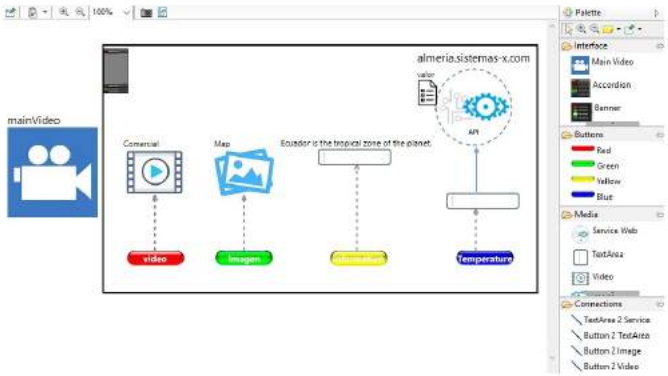
## 5. Concluciones y trabajos futuros

Este documento propone (1) una arquitectura para dispositivos Cross-device empleando técnicas de modelado para describir y respaldar el desarrollo de aplicaciones IoT, y (2) una metodología para el desarrollo automático de aplicaciones para DTV y nodos de hardware. Para validar la propuesta, hemos probado el uso del teléfono inteligente, la televisión digital y un nodo de hardware. De esta manera, se han diseñado aplicaciones en distintos ámbitos del IoT como es T-Health y Smart Home. Estos ámbitos de aplicación han sido considerados porque son de gran importancia para las personas.

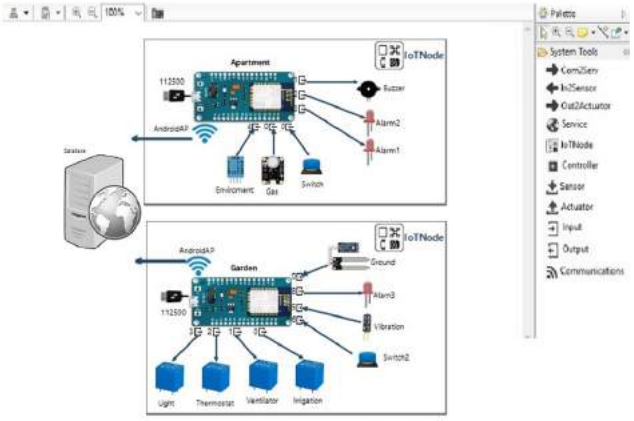
Como trabajo futuro, algunas segundas líneas aún están abiertas: a) ampliación de la metodología para la generación automática de código para las plataformas móviles y web, b) ampliación del metamodelo para incorporar una mayor versatilidad en el diseño de las interfaces, y c) pruebas de rendimiento.




a) Diagrama de una arquitectura para un Parking.




b) Diagrama de una aplicacion de Metrología para DTV.




c) Diagrama de una aplicación de Smart Home.

**Figura 4.** Escenarios de aplicación.

## Agradecimientos

Este trabajo ha sido financiado por el MINECO en el marco de los proyectos TIN2013-41576-R y TIN2017-83964-R.

## Referencias

1. Ammar, M., Russello, G. and Crispo, B.: Internet of Things: A survey on the security of IoT frameworks. Journal of Information Security and Applications. Elsevier Ltd, 38, pp. 8–27. (2018)
2. Alves, G., Barbosa, R., Kulesza, R., and Filho, G.: A Software Testing Process for Ginga Products. Applications and Usability of Interactive TV. CCIS, Vol. 689, pp 61–73. Springer (2016).
3. Atzori, L., Iera, A. and Morabito, G.: From "Smart Objects"to "Social Objects": The Next Evolutionary Step of the IoT. IEEE Com. Mag., pp. 97-105. (2014)
4. Bailey, J., Budgen, D., Turner, M., Kitchenham, B., Brereton, P., and Linkman, S.: How Software Designs Decay: A Pilot Study of Pattern Evolution, Proceedings - 1st International Symposium on Empirical Software Engineering and Measurement. ESEM 2007. pp. 449–51. (2007)
5. Benouda, H., Azizi, M., Esbai, R., Moussaoui, M.: MDA Approach to Automate Code Generation for Mobile Application. Mobile and Wireless Technologies, Springer, pp. 241–250. (2016)
6. Chmielewski, J.: Device-Independent Architecture for Ubiquitous Applications. Pers Ubiquit Comput. (18):481–488. Springer (2014)
7. Criado, J., Asencio, J., Padilla, N., and Iribarne, L.: Integrating Cyber-Physical Systems in a Component-Based Approach for Smart Homes. Sensors. 18(7),2156. (2018)
8. García, C.G., Espada, J.P., Núñez-Valdez, E.R., García-Díaz, V.: Midgar: Domain-specific language to generate smart objects for an internet of things platform. 8th Conf. on Inn. Mobile & Internet Services in Ubiq. Comp., pp. 352–357. (2014)
9. Gomes, T., Lopes, P., Alves, J., Mestre, P., Cabral, j., Monteiro, and Tavares, A.: A Modeling Domain-Specific Language for IoT-Enabled Operating Systems. Annual Conf. of the IEEE Industrial Electronics Society pp. 3945–3950. IEEE (2017)
10. Gonçalo, M., Garcia N. and Pombo N.: A Survey on IoT: Architectures, Elements, Applications, QoS, Platforms and Security Concepts. Adv. in Mobile Cloud Computing and Big Data in the 5G Era. 22. Springer (2016)
11. Malavolta, I., Mostarda, L., Muccini, H., Ever, E., Doddapaneni, K. and Gemikonakli, O.: A4WSN: An Architecture Driven Modelling Platform for Analysing and Developing WSNs. Software and Systems Modeling. pp. 1-21. Springer (2018)
12. Ribeiro, A. and Rodrigues, A.: Evaluation of XIS-Mobile, a Domain Specific Language for Mobile Application Development. Journal of Software Engineering and Applications, 7(11), pp. 906–919. (2014)
13. Rodríguez, J., Cueva, J., Montenegro, C., Granados, J., González, R.: Metamodel for Integration of Internet of Things, Social Networks, the Cloud and Industry 4.0. J. of Ambient Intell. and Humanized Computing 9(3):709–23. Springer (2017)
14. Schachinger, D., and Kastner,k.: Model-driven integration of building automation systems into Web service gateways. Proceedings - IEEE World Conference on Factory Communication Systems, WFCS 2015,pp. 1-6 IEEE (2015)

15. Troya, J., Vallecillo, A. Durán, F., Zschaler, S.: Model-driven performance analysis of rule-based domain specific visual models, Information and Software Technology. (2013)
16. Yao, L., Sheng, Q., Benatallah, B. Dustdar, S., Wang, X., Shemshadi, A. and Kanhere, S.: WITS: An IoT Endowed Computational Framework for Activity Recognition in Personalized Smart Homes.(4):369–85. Springer (2018)

# Detección de una matriz copositiva mediante la evaluación de las facetas de un simplex unidad

Jose Manuel García Salmerón

Grupo Supercomputación-Algoritmos, Departamento de Informática, Universidad de Almería, josemanuel@ual.es

**Resumen** El problema de encontrar el mínimo de un problema de optimización cuadrática estándar permite determinar cuando la matriz simétrica involucrada en la formulación es copositiva. Recientemente se ha desarrollado un algoritmo que evalúa las facetas de un simplex unidad para determinar si la función cuadrática es positiva en una faceta. Se desarrollaron diferentes tests para descartar facetas de la búsqueda donde se puede verificar que la función cuadrática es positiva o para reducir la dimensión de una faceta al ser la función cuadrática monótona en ella. Aunque los test mencionados permiten reducir el número de facetas a evaluar, este número puede ser muy elevado para algunas matrices copositivas. Debido a que la evaluación de una faceta puede realizase de forma independiente de la evaluación de otra faceta, este problema es un buen candidato a ser resuelto de forma paralela. Aún así, hay que tener en cuenta que varias facetas comparten sub-facetas que no deben ser evaluadas si se probó que la función cuadrática es positiva en al menos una de las facetas padre. En este estudio se muestra una posible implementación paralela teniendo también en cuenta un uso eficiente de la memoria.

**Keywords:** Matriz copositiva, simplex unidad, faceta, memoria compartida, paralelismo

## 1. Introducción

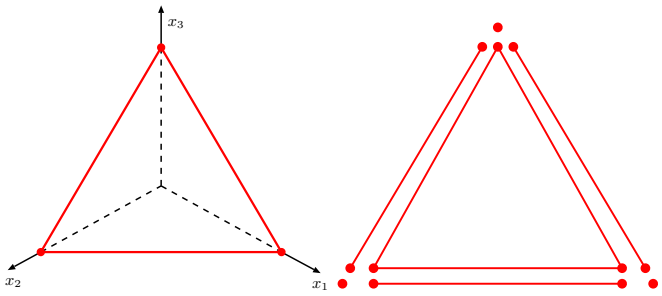
La copositividad juega un papel importante en la optimización combinatoria y cuadrática. Si se establece un problema de optimización lineal sobre el cono copositivo se obtienen reformulaciones exactas de problemas combinatorios. Como ejemplo, podemos mostrar el problema del Clique máximo, que es un problema NP-Completo [1]. Otras aplicaciones interesantes conectadas a este problema son las tratadas por [2, 4–6]. Sea $\mathbf{1}$ el vector con todos los elementos a uno en la dimensión apropiada, $\omega(G)$ el número Clique de un grafo $G$, $I = \mathbf{1}\mathbf{1}^T$ la matriz con todos sus elementos iguales a uno, $t \in \mathbb{N}$ un escalar, y $Q = I - A_G$ una matriz obtenida de la matriz de adyacencia $A_G$ del grafo $G$. El objetivo de este problema de programación copositiva es encontrar el menor valor de $t$ tal que $tQ - I$ es copositiva, i.e., está en el conjunto $\mathcal{C}$ de las matrices copositivas,

$$\omega(G) = \text{mín}\{t : \ tQ - I \in \mathcal{C}\}. \tag{1}$$

Una matriz simétrica $A$ de $n \times n$ se certifica que no es copositiva cuando $\exists x \in S_n,\ x^T Ax < 0$, en el simplex unidad

$$S_n = \{x \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i = 1;\ x_i \geq 0, \quad i = 1, \ldots, n\}. \tag{2}$$

Se define una faceta $F_k \subseteq S_n$ como un subconjunto de los vectores unitarios $e_i$ que están identificados por el vector binario $z \in \{0,1\}^n$, donde $z_i = 1$ si $e_i$ está incluido en el conjunto de vértices. El número de vértices en la faceta está determinado por $m = \sum_i z_i$, el cual también determina su número de facetas (incluyéndose ella misma y los vértices) $k = 2^n - 1$. Por lo tanto, $z = \mathbf{1} = (2^n - 1)_2$ se refiere al simplex unidad indexado como $F_{2^n-1}$. Una faceta $F_k$ es en realidad un símplice unidad de dimensión $m$ ($S_m$) cuyas componentes están determinadas por $z$ $\left(m = \sum_i z_i,\ z = (k)_2\right)$. La figura 1a) muestra un simplex unidad de dimensión 3. La figura 1b) intenta mostrar sus facetas: el propio simplex, tres lados y tres puntos. Los conjuntos de cuatro puntos más cercanos son en realidad el mismo punto en el espacio, donde solo una componente vale uno y las demás valen cero.




(a) Simplex unidad 3D. (b) Facetas simplex 3D.

Figura 1: Simplex unidad y facetas 3D.

Además de la matriz identidad $E_m = (e_1, \ldots, e_m)$ en el espacio $m$ dimensional, se usará también la matriz $D_m = E_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T$ con las posibles direcciones sobre las facetas del simplex unidad.

Un problema relacionado es el de la programación cuadrática estándar (StQP)

$$f^* := \min_{x \in S_n} f(x) := x^T Ax, \tag{3}$$

donde si $f^* < 0$ entonces $A$ no es copositiva y si $f^* \geq 0$ entonces $A$ es copositiva.

Existen algoritmos, como los de ramificación y acotación [3], que permiten certificar que una matriz no es copositiva o que es $\epsilon$-copositiva. Mediante el refinamiento de símplices, la certificación de la $\epsilon$-copositividad requiere mucha más computación que verificar que una matriz no es copositiva. En [8], donde se usó computación grid, se mostró la verificación de matrices no copositivas con una dimensión de hasta varios miles. Sin embargo, la certificación $\epsilon$-copositiva

de una matriz solo pudo realizarse en un tiempo razonable para dimensiones de hasta $n = 22$.

Se sabe que el óptimo de StQP para una matriz no-positiva semi-definida se encuentra en una faceta. Sin embargo, los algoritmos de ramificación y acotación espaciales antes mencionados evalúan puntos en el interior relativo de $S_n$. Por ello se desarrolló un algoritmo que busca de forma sistemática el punto con el menor valor de la función cuadrática en cada una de las $2^n - 1$ facetas.

Este trabajo tiene la siguiente organización. En la sección 2 se describen los desarrollos matemáticos para determinar cuando una matriz es positiva semi-definida y la función cuadrática (3) es monótona en una faceta. La sección 3 muestra una primera versión secuencial del algoritmo y sus desventajas, además de una versión mejorada basada en niveles. La sección 4 presenta una posible versión paralela de la versión secuencial basada en niveles. Finalmente la sección 5 muestra los resultados del algoritmo secuencial y paralelo señalando las principales conclusiones.

## 2. Caracterización del mínimo de un StQP

Para evaluar $f$ en la faceta $F_k$ hay que considerar la matriz $A_k$.

**Definición 1.** *Dada una matriz $A$ simétrica de $n \times n$ y un vector binario $z$, $A_k$ es una matriz de $m \times m$ con $m < n$ y las filas, columnas $i$ de $A$, seleccionadas si $z_i = 1$.*

Por lo tanto, la optimización de $\min_{x \in F_k} x^T Ax$ en una faceta es equivalente a $\min_{x \in S_m} x^T A_k x$.

### 2.1. Caracterización del óptimo relativo interior

Un resultado de la programación cuadrática muestra que el mínimo de una función cuadrática indefinida puede encontrarse en el límite del espacio factible [7]. Esto es interesante para el problema (3) ya que el mínimo no se encuentra en el interior de $S_n$, aunque si puede estar en el interior de una de sus facetas. Se define el interior de $\breve{S}_m$ de $S_m$ como

$$\breve{S}_m = \Big\{ x \in \mathbb{R}^m \mid \sum_{j=1}^{m} x_j = 1;\ x_j > 0, \quad j = 1, \ldots, m \Big\}.$$

**Proposición 1.** *Si $\exists x^* \in \operatorname{argmin}_{S_n} f(x) \in \breve{S}_n$, entonces $D_n A x^* = 0$ y $H = D_n A D_n$ es una matriz semi-definida.*

La proposición 1 puede aplicarse a cualquier faceta $F_k \subset S_n$.

**Corolario 1.** *Si $\exists x^* \in \operatorname{argmin}_{F_k} f(x) \in \breve{F}_k$, entonces $\exists y^* \in \breve{S}_m$, $D_m A_k y^* = 0$ y $D_m A_k D_m$ es positiva semi-definida.*

Consideremos la matriz de vértices $V$ con su correspondiente cobertura convexa (símplice $\Delta$)

$$\begin{aligned} \Delta =& \{x = V\lambda, \sum \lambda_i = 1, \lambda_i \geq 0, i = 1, \ldots, n\} \\ \Delta =& \{x = V\lambda, \lambda \in S_n\} \end{aligned} \tag{4}$$

y $Q = V^T AV$, tal que $f(x) = x^T Ax$ corresponde con $\lambda^T V^T AV\lambda = \lambda^T Q\lambda$. Usando está notación, el Corolario 1 se transforma en $DQ\lambda^* = DVAx^* = 0$. Aunque es complicado, este análisis puede extenderse a las facetas de menor dimensión. Sin embargo, la cuestión es ¿en qué faceta se encuentra el mínimo? Para ello, hay que tener en cuenta las derivadas direccionales $DAx$ y $DQ\lambda$.

### 2.2. Consideraciones sobre monotonía

**Proposición 2.** *Sea $V$ una matriz de vértices, $\Delta$ de la ecuación (4), la matriz $D$ con columnas $d_i$ y $Q = V^T AV$. Si $d_i^T Q \geq 0$, entonces $\exists x^* \in \operatorname{argmin}_\Delta f(x)$ en la faceta que corresponde con el valor de $\lambda_i = 0$.*

La importancia de este resultado es que, para el problema (3), si se cumple la condición de la proposición 2, se puede reducir la investigación de la copositividad de $A$ a una faceta $F_k$ de $S_n$ que tiene una menor dimensión. Por lo tanto, podemos buscar un valor negativo de $f$ en el área más prometedora de $S_n$ reduciendo $S_n$ a $F_k$. Cuando $d_i^T Q \geq 0$, la faceta $F_k$ está definida por una matriz $\hat{V}$ de vértices de $m \times n$ que se obtiene quitando el vértice $v_i$ de $V$. Entonces $\hat{Q} = \hat{V}^T A\hat{V}$. Esto se puede extender de forma directa a facetas de menor dimensión que si cumplen la Proposición 2 usando $\hat{Q}$ y el correspondiente $D_m$, se pueden reducir a una de sus sub-facetas.

## 3. Algoritmo secuencial

El Algoritmo 1 enumera las facetas $F_k$ de $S_n$, marcando como *Chequeadas* aquellas donde no puede estar el mínimo de StQP (3) o el mínimo en la faceta $F_k$ es positivo, hasta que se haya chequeado la lista completa de facetas o se haya encontrado un punto cuyo valor es negativo. Si la matriz $A$ no contiene valores negativos entonces es copositiva (línea 1). Esto también ocurre con las sub-facetas $F_k$, donde si $A_k$ no contiene valores positivos, el optimo interior de $F_k$ es positivo y las sub-facetas de $F_k$ se marcan como chequeadas (línea 9).

Se usa la Proposición 2 para chequear si $f$ es monótona creciente en una de las componentes de la faceta (línea 12) y en caso afirmativo se reduce a una de sus sub-facetas de dimensión menor. Si no es el caso, se evalúa si $H_k$ es positiva semi-definida (línea 15) condición necesaria para tener un óptimo interior. Para que una matriz sea positiva semi-definida sus autovalores deben ser positivos. Si $H_k$ no es es positiva semi-definida, la solución de StQP (3) solo puede estar en sus sub-facetas. En el caso de que $H_k$ sea positiva semi-definida, hay que encontrar el punto estacionario relativo interior de $F_k$ que resuelva $D_m A_k x = 0$

**Algoritmo 1** Test de copositividad basado en facetas

**Entrada:** $A$: matriz simétrica de $n \times n$.

```
 1: if A ≥ 0 then
 2:    return A es copositiva
 3: if A_{i,i}, i = 1,...,n, o f(1/n 1) es negativo then
 4:    return A no es copositiva
 5: k = 2^n − 1
 6: Marcar las facetas F_i, i = k,...,n como no chequeadas.
 7: while k > 2 do
 8:    if F_k no chequeada then
 9:       if A_k ≥ 0 then
10:          Marcar las sub-facetas de F_k como chequeadas
11:          break
12:       if ∃i, D_{mi}^T A_k > 0, i.e. monótona creciente en la dirección i then
13:          Marcar las sub-facetas que no tengan x_i = 0 como chequeadas
14:       else
15:          if H_k es positiva semi-definida then
16:             if ∃x* ∈ S_m : D_m A_k x* = 0 then
17:                if x*^T A_k x* < 0 then
18:                   return A no es copositiva
19:                else
20:                   Marcar las sub-facetas de F_k como chequeadas
21:    k = k − 1
22: return A es copositiva
```

en $S_m$ (línea 16). Para ello hay que resolver un problema de programación lineal que maximiza el valor del elemento mínimo, de forma que se tiene una solución en $S_m$ o es cierto que no existe una solución de

$$\begin{aligned} &\text{máx}\, g, \\ &\text{s.t: } D_m A_k x = 0, \\ &\quad \mathbf{1}^T x = 1, \\ &\quad g \leq x_i,\ i = 1, \ldots, m. \end{aligned} \tag{5}$$

Si el resultado de (5) es $g < 0$, entonces no existe una solución en $S_m$ y la solución de StQP (3) no está ni en $F_k$ ni en sus sub-facetas.

El algoritmo 1 presenta dos principales inconvenientes: i) las facetas no se visitan de mayor a menor dimensión y ii) la lista completa de facetas chequeadas puede llegar a consumir mucha memoria.

En cuanto al primer inconveniente, la Figura 2 muestra el conjunto de facetas para $n$=4. Si se visitan en orden descendiente, se evaluaría antes $F_{12}$ que tiene una dimension menor que $F_7$, que en caso de ser descartada junto con sus sub-facetas ($F_3$, $F_5$, $F_6$) y las sub-facetas de estas, descartarían más facetas del árbol que en el caso de descartar $F_{12}$. El número de niveles del árbol es $n$, estando $S_n$ en el nivel $n$. Cada nivel tiene $\binom{n}{\text{nivel}}$ elementos. El nivel indica el número de bits igual a uno en los índices binarios de las facetas de ese nivel que, como se

indicó anteriormente, se corresponde con las coordenadas que pueden tomar un valor en el rango [0,1].

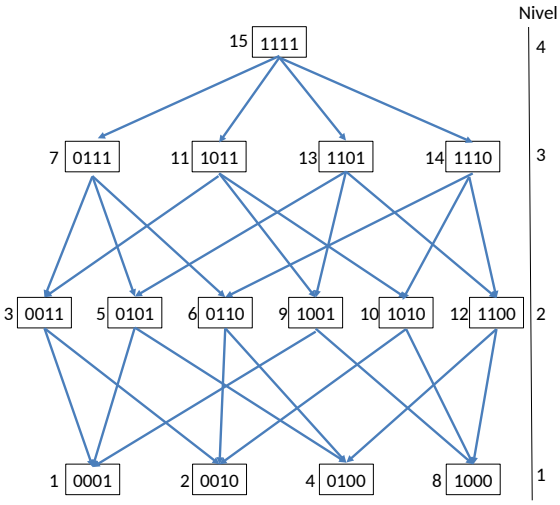



Figura 2: Árbol de facetas para $n = 4$.

En cuanto al segundo inconveniente podemos mostrar un ejemplo. Para $n = 28$, si se usa un `char` para almacenar si la faceta $F_k$ ha sido chequeada o no, se necesitarían $2^n - 1$ bytes.

Estos inconvenientes hacen pensar que es mejor trabajar con solo dos niveles, el actual y el siguiente. Las facetas del siguiente nivel se marcarían como chequeadas (True) dependiendo de las facetas del nivel actual. Por lo tanto se usarán dos vectores de booleanos: NA (Nivel Actual) y NS (Nivel Siguiente) con $\binom{n}{nivel}$ y $\binom{n}{nivel-1}$ elementos respectivamente, donde *nivel* es el nivel actual. Se necesitarán las funciones $k$=IndexFaceta($n$, nivel, i) que devuelve el indice de la faceta en la posición $i$ del nivel, e $i$=PosNivelFaceta($n$,nivel,$k$) que devuelve la posición de la faceta $k$ en el vector de su nivel. Por ejemplo, en la Figura 2, 5=PosNivelFaceta(4,2,10) y 10=IndexFaceta(4,2,5). Como puede observarse en la Figura 2 cada faceta del nivel $j$ tiene $j$ sub-facetas en el nivel $j - 1$.

Podría pensarse que usar dos vectores de `char` no reduce mucho la memoria necesaria ya que los vectores actual y siguiente de mayor tamaño se encontrarían en los niveles de la mitad del árbol. En la Figura 2, entre los niveles 3 y 2 ocupan 10 de los 15 `char` totales, pero esta diferencia se agranda conforme lo hace $n$.

El algoritmo 2 muestra el pseudocódigo para chequear la copositividad de una matriz simétrica basado en facetas del simplex unidad por niveles.

## 4. Versión paralela

En esta sección se estudia la versión paralela del algoritmo 2. El esquema se basa en un modelo maestro-esclavo donde el maestro reparte aquellas facetas del nivel actual que deben evaluarse entre los distintos esclavos. Un esclavo recibirá

**Algoritmo 2** Test de copositividad basado en facetas por niveles

**Entrada:** $A$: matriz simétrica de $n \times n$.

**if** $A \geq 0$ **then**
  **return** A es copositiva
**if** $A_{i,i},\ i = 1, \ldots, n$, o $f(\frac{1}{n}\mathbf{1})$ es negativo **then**
  **return** A no es copositiva
$\text{NA}_1$=False ▶ Nivel Actual. False=No Chequeada
$\text{NS}_i$=False, $i = 1, \ldots, n$. ▶ Nivel Siguiente
**for** nivel=$n$ to 2 **do**
  **for** $i$=1 to $\binom{n}{\text{nivel}}$ **do**
    k=IndexFaceta($n$,nivel,$i$)
    **if** $\text{NA}_i$==True or $A_k \geq 0$ **then**
      Marcar las sub-facetas de $F_k$ en NS como True.
    **else**
      **if** $\exists i$, $D_{mi}^T A_k > 0$, i.e. monótona creciente en la dirección $i$ **then**
        Marcar las sub-facetas de $F_k$ que no tengan $x_i = 0$ en NS como True.
      **else**
        **if** $H_k$ es positiva semi-definida **then**
          **if** $\exists x^* \in S_m : D_m A_k x^* = 0$ **then**
            **if** $x^{*T} A_k x^* < 0$ **then**
              **return** A no es copositiva
            **else**
              Marcar las sub-facetas de $F_k$ en NS como True.
  **if** nivel>2 **then**
    NA=NS
    $\text{NS}_i$=False, $i = 1, \ldots, \binom{n}{\text{nivel}-2}$
**return** A es copositiva

del maestro el (los) índice(s) del vector NA (Nivel Actual) marcados con False y evaluará las facetas correspondientes, devolviendo los índices del vector NS (Nivel Siguiente) que deben establecerse a True (Chequeados).

La principal limitación de la versión paralela son las barreras que existen en la terminación de cada nivel, ya que no se debería empezar a procesar el siguiente nivel sin haber terminado el anterior, si se quiere evitar evaluar una faceta que podría no ser necesario tener que evaluarla en el nivel siguiente.

Otra limitación del paralelismo es que el maestro tras recibir los índices del siguiente nivel (NS) de los esclavos que deben marcarse a True, los tiene que marcar y esta operación se hace de forma secuencial.

Existe un compromiso entre cuantos indices del nivel actual se le dan a cada esclavo y el número de escrituras secuenciales en el nivel siguiente que debe realizar el maestro. Mientras que enviar un solo indice al esclavo mejora el balanceo de la carga entre esclavos, enviar más de un indice hace que el esclavo elimine los índices repetidos, reduciendo así el número de escrituras secuenciales del maestro. El maestro también debe marcar los indices del siguiente nivel que correspondan a sub-facetas de la faceta marcada como chequeada en el nivel actual.

Resumiendo, el trabajo del esclavo es evaluar una(s) faceta(s) y devolver si ha encontrado un punto negativo o los indices de las facetas del siguiente nivel a marcar como chequeadas. El trabajo del maestro es repartir las facetas a evaluar entre los esclavos y marcar los índices del siguiente nivel como chequeados (True) para las facetas del nivel actual que estén a True y los indices devueltos por los esclavos. Como el maestro es el único que escribe en la lista del siguiente nivel no se necesita exclusión mutua. Se podría pensar en un acceso paralelo a la lista del siguiente nivel ya que solo se realizan escrituras pero como muestra la Figura 2 el acceso a una faceta del nivel inferior desde otra del nivel superior no es uniforme por lo que es difícil mantener una localidad espacial de los datos.

Como se ha dicho anteriormente mandar más de una faceta al esclavo permite que el maestro realice menos escrituras en el nivel siguiente. Otra posible estrategia es partir el vector del nivel actual en trozos. Cada trozo se le envía a un esclavo siguiendo un modelo parecido al usado en el algoritmo paralelo anterior, ademas del indice de inicio del trozo en el vector del nivel actual. El trozo de vector recibido por el esclavo puede contener facetas chequeadas y sin chequear. Si la faceta está chequeada, el esclavo genera los indices de las sub-facetas en el nivel siguiente a marcar como chequeadas y si no está chequeada la evalúa. La evaluación de una faceta por parte del esclavo puede dar como resultado un punto negativo o una lista de sub-facetas en el siguiente nivel que se unirá a la lista ya existente de otras facetas chequeadas o evaluadas anteriormente, eliminando los posibles duplicados. Esa lista se enviará finalmente al maestro. En este esquema los esclavos descargan de trabajo al maestro ya que ahora el maestro solo trocea el vector del nivel actual y actualiza los indices del nivel siguiente a partir de las listas recibidas de los esclavos.

## 5. Resultados y conclusiones

En está sección se mostraran algunos de los casos evaluados por el algoritmo paralelo planteado en la sección anterior, comparando su rendimiento con el algoritmo desarrollado para computación grid en [8]. La diferencia principal de ambos algoritmos radica en la forma en que refinan el espacio, mientras que en [8] optan por un refinamiento del espacio mediante bisección, en este trabajo hemos desarrollado un método de refinamiento basado en descomposición del simplex en facetas de menor dimensión. El algoritmo paralelo ha sido desarrollado en Matlab y ha sido ejecutado para las pruebas en la maquina Bullion S8, que dispone de 8 Intel Xeon 8860v3 (19 cores) y 2,3 TB de memoria RAM.

La Figura 3 muestra la aceleración del algoritmo paralelo para el caso de un problema de dimensión 16 en Bullion. En la gráfica se observa como el resultado obtenido (P5, linea azul) está bastante alejado del resultado ideal (Aceleración lineal, linea negra), sobre todo para los casos con más de 8 hebras. Esto hace indicar que el algoritmo paralelo no está totalmente optimizado.

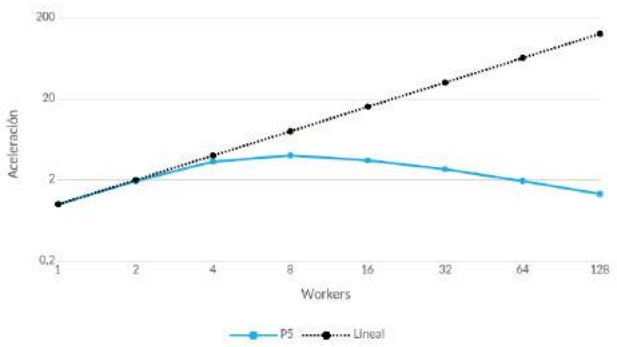



Figura 3: Aceleración del problema P5 de 16 dimensiones en Bullion.

| Hebras | Tiempo (seg) |
|---|---|
| 1 | 58,74 |
| 2 | 30,19 |
| 4 | 17,49 |
| 8 | 14,63 |
| 16 | 16,84 |
| 32 | 21,71 |
| 64 | 30,19 |
| 128 | 43,67 |

Tabla 1: Tiempo por hebras ejecutadas del problema P5 de 16 dimensiones en Bullion.

Si analizamos con algunos datos más esté problema podemos descubrir las razones de este rendimiento.

La Tabla 1 muestra el tiempo de las ejecuciones del problema para las diferentes cantidades de hebras empleadas. El tiempo decrece hasta alcanzar las 16 hebras, en las que la aceleración positiva llega a su fin, probablemente debido a la poca cantidad de trabajo de este problema o al desbalanceo del mismo. Hay que recordar que existen barreras de sincronización entre niveles.

La Tabla 2 muestra el trabajo en cada nivel y se observa que solo en niveles interiores es posible aprovechar el paralelimo debido a una suficiente cantidad de trabajo. Todo ello, sumado al tipo de memoria de la maquina Bullion (NUMA), puede provocar un bajo aprovechamiento del trabajo en paralelo en los nodos.

En cualquier caso, y pese a los desanlentadores primeros resultados, si comparamos el tiempo de ejecución del algoritmo basado en facetas con el del algoritmo basado en bisección los resultados no son tan malos. El tiempo para el algoritmo secuencial que emplea bisección para el problema P5 de 16 dimensiones es de más de 6 días, mientras que nuestro algoritmo resuelve el mismo problema en poco menos de 1 minuto.

| Nivel (Dim. Facetas) | Nodos |
|---|---|
| 16 | 1 |
| 15 | 16 |
| 14 | 120 |
| 13 | 560 |
| 12 | 1.820 |
| 11 | 4.368 |
| 10 | 8.008 |
| 9 | 11.401 |
| 8 | 12.699 |
| 7 | 10.997 |
| 6 | 7.261 |
| 5 | 3.031 |
| 4 | 477 |
| 3 | 22 |
| 2 | 0 |

Tabla 2: Nodos por nivel del problema P5 de 16 dimensiones en Bullion.

Continuando con la comparación entre ambas metodologías de refinamiento del espacio de búsqueda, si buscamos un problema de mayor tamaño resuelto mediante bisección podemos encontrar uno de 22 dimensiones que necesito de computación grid y de 5,6 horas, y lo comparamos con los resultados de la Tabla 3 en el que empleando 128 hebras y 3 horas pudimos resolver un problema de 28

dimensiones los resultados iniciales ya no parecen tan malos, todo lo contrario, nos alientan a seguir trabajando para optimizar el algoritmo y ver hasta donde podemos llegar.

| Hebras | Tiempo (horas) |
|---|---|
| 32 | 25 |
| 64 | 7 |
| 128 | 3 |

Tabla 3: Tiempo por hebras ejecutadas del problema 28 dimensiones en Bullion.

## Referencias


1. I. M. Bomze, M. Dür, E. de Klerk, C. Roos, A. J. Quist, and T. Terlaky. On copositive programming and standard quadratic optimization problems. *Journal of Global Optimization*, 18(4):301–320, 2000.
2. Immanuel M. Bomze, Werner Schachinger, and Gabriele Uchida. Think co(mpletely)-positive! matrix properties, examples and a clustered bibliography on copositive optimization. *Journal of Global Optimization*, 3(52):423–445, 2012.
3. S. Bundfuss and M. Dür. Algorithmic copositivity detection by simplicial partition. *Linear Algebra and its Applications*, 428(7):1511–1523, 2008.
4. Peter J.C. Dickinson. The copositive cone, the completely positive cone and their generalisations. *PhD thesis, University of Groningen*, 2013.
5. Mirjam Dür. Copositive programming - a survey. *Springer Berlin Heidelberg*, pages 3–20, 2010.
6. Jean-Baptiste Hiriart-Urruty and Alberto Seeger. A variational approach to copositive matrices. *SIAM Review*, 4(52):593–629, 2010.
7. Reiner Horst and Hoang Tuy. *Global Optimization. Deterministic Approaches.* Springer, 3rd edition, 1996.
8. J. Žilinskas and M. Dür. Depth-first simplicial partition for copositivity detection, with an application to maxclique. *Optimization Methods and Software*, 26(3):499–510, 2011.

# Avances en el Modelado y Simulación de un Nuevo Concepto de Vehículo Urbano Eléctrico Ligero. Almacenamiento y Distribución de Energía.

Francisco José Gómez Navarro

Universidad de Almería
Carretera Sacramento s/n
04120 La Cañada de San Urbano
Almería - España
https://www.ual.es/

**Resumen** El presente trabajo tiene por objeto, dar a conocer el avance del proyecto de investigación sobre un nuevo concepto de vehículo eléctrico ligero, que aprovecha las diversas fuentes de energía renovable disponibles, optimizando su uso, al objeto de maximizar la reducción en la emisión de gases de efecto invernadero ligados al transporte urbano de personas y bienes. Durante este periodo, con la ayuda del lenguaje de modelado Modelica® [1] y la herramienta de modelado Dymola® [2], se han desarrollado los modelos correspondientes al almacenamiento y distribución de energía del vehículo. Se ha desarrollado y validado un modelo dinámico que simula adecuadamente el comportamiento de baterías del tipo Litio-Ferrofosfato ($LiFePO_4$) y que supone una aportación significativa respecto a los modelos de baterías Ion-Litio ($Li-ion$) encontrados en la bibliografía y que son de uso común en los vehículos eléctricos. También se ha trabajado en el desarrollo de un modelo dinámico para un convertidor CC/CC bidireccional que permita la conexión de los distintos elementos generadores y consumidores de energía del vehículo a un bus de energía común.

**Keywords:** Vehículo Eléctrico, Movilidad Urbana, Energías Renovables, Reducción CO2, Modelado Orientado a Objetos, Modelica, Batería

## 1. Introducción

La sociedad moderna ha basado su desarrollo en gran medida en la posibilidad de desplazar cantidades suficientes de bienes y personas entre distintas localizaciones de forma eficaz. El transporte consume el 19 % de la energía a nivel mundial y emite el 23 % del dióxido de carbono ($CO_2$) debido al consumo energético [3]. Con la tendencia actual, el uso de energía para el transporte aumentará un 50 % para 2030 y más de un 80 % para 2050.
El Intergovernmental Panel on Climate Change (IPCC) advierte que para evitar las desastrosas consecuencias del cambio climático, las emisiones globales de

$CO_2$ deben disminuir, al menos, un 50 % de aquí al año 2050 [4]. El transporte juega un papel decisivo en la consecución de este objetivo, resulta indispensable por tanto incidir en la necesaria adaptación de los modos actuales de desplazamiento. En Europa, se han establecido las bases para una política de transportes competitiva y sostenible[5], entre las que cabe destacar:

- La eliminación progresiva de los vehículos de «propulsión convencional» en el entorno urbano es una contribución fundamental a una reducción significativa de la dependencia del petróleo, las emisiones de gases de efecto invernadero, la contaminación atmosférica local y la contaminación acústica.
- Debe fomentarse el uso de vehículos de pasajeros más pequeños, más ligeros y más especializados en el transporte por carretera.

El transporte ligero, fundamentalmente orientado a personas, consumió en 2006 el 47 % de la energía dedicada al transporte [3]. El parque mundial de vehículos ligeros (Ligh Duty Vehicle - LDV) es previsible que se triplique para el año 2050, principalmente debido al incremento en países en vías de desarrollo[6]. De lo expuesto anteriormente, se puede concluir que uno de los focos de actuación preferentes para la consecución del objetivo de reducción de la emisión de gases de efecto invernadero está en el **transporte urbano ligero** y en el uso de **fuentes de energía alternativas a los combustibles fósiles**. Estas fuentes deben:

- Ser renovables para evitar su agotamiento.
- Estar disponibles en la zona para evitar la dependencia energética de terceros.
- Ser acumulables para poder disponer de las reservas adecuadas que equilibren la capacidad de producción y la demanda.

### 1.1. Motivación

El trabajo de investigación propuesto viene justificado por dos motivaciones diferentes y complementarias:

- Necesidad de disponer de modelos adecuados que permitan analizar y anticipar las prestaciones y comportamientos de las distintas alternativas tecnológicas en estudio para el caso de los sistemas de transporte urbano del futuro. La necesidad de disponer de sistemas de propulsión y fuentes energéticas alternativas obliga a disponer de modelos modulares, con capacidad para integrar y simular el comportamiento dinámico del sistema completo, integrando las distintas tecnologías y desde distintos niveles de abstracción. Si bien hay un elevado número de investigadores que han desarrollado e investigado sobre modelos que analizan las posibilidades de estas nuevas tecnologías de alimentación y propulsión, en la mayoría de los casos se basan en la simple adaptación de la concepción tradicional del vehículo a esta nueva realidad.

- Necesidad de avanzar en nuevas propuestas de movilidad urbana, basadas en vehículos más ligeros, mas adaptados al uso especifico para transporte ligero de corta distancia, con máximo aprovechamiento de las fuentes de energía renovable disponibles y con esquemas de uso que optimicen su capacidad.

### 1.2. Hipótesis

La hipótesis principal que se pretende demostrar, es la viabilidad de un nuevo concepto de movilidad urbana basado en el uso intensivo de fuentes de energía renovable, al que podremos llamar Very Light Urban Vehicle (VLUV). Como hipótesis secundaria, se pretende demostrar cómo este concepto consigue una reducción importante en la emisión de $CO_2$ a la atmósfera y disminuye el consumo de combustibles fósiles y otras fuentes de energía no renovable [7].

## 2. Avances

El primer paso del trabajo de investigación, tras la revisión bibliográfica, ha consistido en la determinación de un concepto genérico de vehículo ligero para desplazamiento urbano que permita la investigación y desarrollo de los modelos correspondientes a cada uno de los componentes, sus interfaces y la simulación del conjunto completo bajo distintas configuraciones y condiciones de contorno. Para el desarrollo de los distintos modelos se ha optado por el Modelado Orientado a Objetos con ayuda del lenguaje de modelado Modelica®. Este estándar abierto tiene las siguientes ventajas desde el punto de vista del modelado y la simulación de sistemas multi-físicos [8][9]:

- Encapsulamiento del conocimiento.
- Capacidad de interconexión topológica.
- Modelado jerárquico.
- Instanciación de objetos.
- Herencia de clases.
- Capacidad de interconexión generalizada.

Se utiliza la herramienta de desarrollo Dymola® [10], basada en el lenguaje de modelado de código abierto Modelica® para el desarrollo, simulación y optimización de los modelos. Para la definición de las clases base, los interfaces de conexionado y la estructuración general del modelo se ha tomado como referencia la librería VehicleInterfaces de Modelica [11]. Esta librería proporciona una serie de definiciones de interface normalizadas para uso en subsistemas de automoción y modelos de vehículos. Su objetivo es el de promover la compatibilidad entre las distintas librerías de componentes de automoción y proporcionar una estructura flexible y potente para el modelado de vehículos.

La estructura final de los modelos componentes del vehículo se pueden apreciar en la figura 1. Con los diferentes modelos desarrollados de cada uno de los distintos componentes, se pueden conseguir las distintas configuraciones deseadas de vehículos que nos ayudarán en la simulación dinámica, análisis y evaluación de resultados.

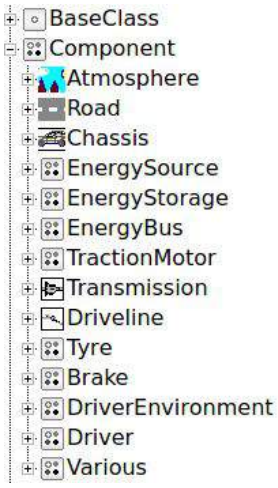



**Figura 1.** Estructura básica de la librería de componentes

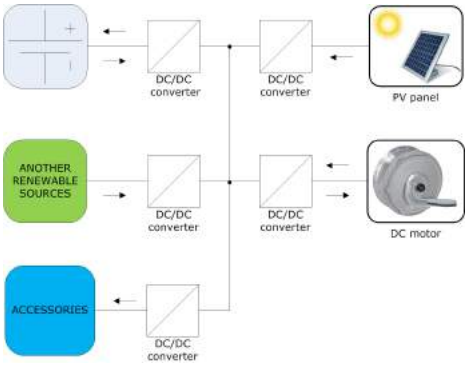



**Figura 2.** Esquema general del clúster inteligente

### 2.1. Modelado del vehículo

Al objeto de la simulación dinámica del vehículo completo, de su control y del análisis del balance energético de su actuación, se ha desarrollado un modelo de clúster de energía inteligente, que agrupa los diferentes elementos componentes del vehículo que aportan o consumen energía. Tal como se aprecia en la figura 2, está compuesto principalmente por el panel fotovoltáico (PV), la batería, los convertidores de tensión DC/DC, el motor de accionamiento del vehículo (que puede funcionar como freno regenerativo), los accesorios del vehículo y otras posibles fuentes de energía renovable en estudio. También se aprecia en la figura los sentidos del flujo de energía posibles en cada uno de los componentes. Los componentes que aportan energía son: la batería (en modo descarga), el panel PV, el motor (en modo freno regenerativo) y las otras fuentes de energía

renovable instaladas. Los componentes que consumen energía son: la batería (en modo carga), el motor de accionamiento del vehículo y los accesorios. Todos los componentes se interconectan mediante un bus DC a través de los convertidores DC/DC funcionando en modo uni ó bidireccional.

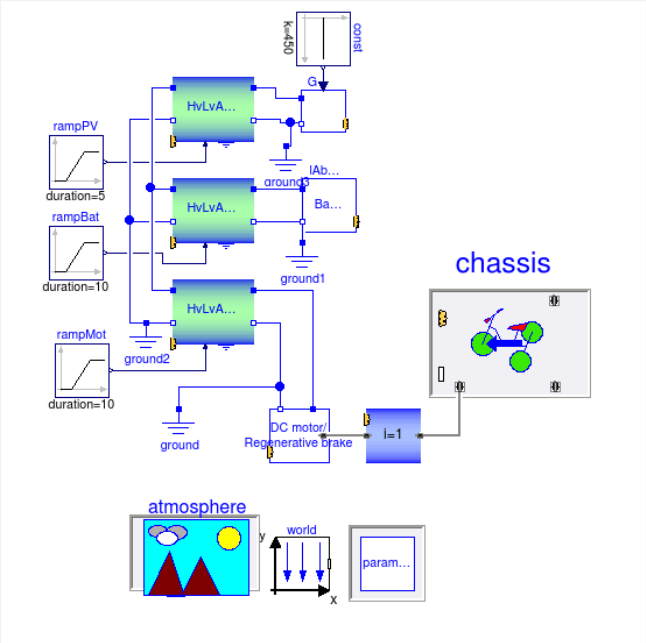



**Figura 3.** Modelo completo del vehículo en Modelica

En la figura 3 se muestra el modelo completo del vehículo en Modelica.

## 2.2. Modelado de la batería

De entre los distintos elementos componentes del vehículo eléctrico, la batería es el que juega un papel más destacado, puesto que de su capacidad y comportamiento dependerán en gran parte las prestaciones finales del vehículo, fundamentalmente en lo relativo a la autonomía, fiabilidad y coste de operación. Resulta fundamental por tanto, disponer de unos modelos de batería suficientemente fiables, que permitan su uso en simulaciones de tie mpo real y que se adapten a las distintas tecnologías disponibles en el mercado, de uso habitual en vehículos eléctricos.
Se ha desarrollado y validado por tanto un modelo dinámico que simula adecuadamente el comportamiento de baterías del tipo $LiFePO_4$ y que supone una aportación significativa respecto a los modelos de baterías $Li-ion$ encontrados en la bibliografía [12]. Este modelo es de aplicación directa a baterías $Li-ion$ y con

pequeñas modificaciones a otras tecnologías habituales como las de plomo-ácido (LA), niquel-hidruro metálico (Ni-MH) o níquel-cadmio (Ni-Cd). Los resultados de esta investigación han sido remitidos para su publicación.

### 2.3. Modelado del convertidor

Tal como podemos ver en la figura 2, cada componente del vehículo que consume o aporta energía al sistema, está unido al mismo mediante un convertidor CC/CC. La función del convertidor es la de adaptar la tensión del bus de contínua del vehículo a la tensión necesaria para el accionamiento de cada componente, en el caso de consumidores, o de adaptar la tensión generada por el componente a la del bus en el caso de generadores. Se está trabajando en el desarrollo de un modelo dinámico de convertidor CC/CC bidireccional que permita la conexión de los distintos elementos generadores y consumidores de energía del vehículo al bus de energía común. Puesto que hay una instanciación múltiple del mismo, es muy importante disponer de un modelo base suficientemente preciso y que permita la simulación en tiempo real del vehículo completo. Sobre este modelo se implementarán los algoritmos de control

## 3. Publicaciones

La siguiente contribución ha sido presentada a congreso internacional:

"Modelling a Smart-Grid for a Solar Powered Electric Vehicle"
Francisco J. Gomez, Luis J. Yebra, Antonio Gimenez
Type of submission: Discussion Contribution
9th International Conference on Mathematical Modelling
Vienna, Feb 21-23, 2018
ISBN 978-3-901608-91-9
DOI: 10.11128/arep.55.a55113

El siguiente artículo ha sido aceptado para publicación, con modificaciones:

"Modelado de Baterías para Aplicacion en Vehículos Urbanos Eléctricos Ligeros"
Francisco J. Gomez, Luis J. Yebra, Antonio Gimenez, José L. Torres
Revista Iberoamericana de Automática e Informática Industrial
Comité Español de automática

## 4. Conclusiones

Se ha realizado la introducción y descripción del proyecto de investigación en curso. Se han revisado las motivaciones y justificaciones del mismo, así como del estado de avance. Los trabajos en curso, una vez definidas las clases base e interfaces a utilizar, están orientados a conseguir el modelo dinámico completo de un

clúster de energía inteligente para ser usado en un vehículo animado por energía solar y otras fuentes de energía renovable, con el propósito del diseño y optimización del control del mismo. El modelo será utilizado para realizar las simulaciones en tiempo real que permitan la optimización de los algoritmos de control de los distintos componentes. Se han presentado los principales componentes, su esquema de interconexión, asi como los flujos de energía. Se ha presentado un ejemplo del modelo completo del vehículo en Modelica. Se han presentado los avances en el desarrollo de un nuevo modelo de batería que permite simular adecuadamente el comportamiento específico de la tecnología $LiFePO_4$ y que es válido para los otros tipos de batería de uso corriente en vehículos eléctricos (LA, Li-ion, Ni-Cd, Ni-MH). También se ha presentado el avance del trabajo de desarrollo del convertidor CC/CC de interconexión entre los distintos componentes.

## Referencias


1. P. Fritzson, *Principles of Object-Oriented Modeling and Simulation with Modelica 3.3: A Cyber-Physical Approach*, 2nd ed. Linköping - Sweden: John Wiley & Sons - IEEE Press, 2015. [Online]. Available: https://books.google.es/books?id=wgIaBgAAQBAJ
2. Dassault Systèmes AB, "Dymola - Dynamic Modeling Laboratory - User Manual," Lund - Sweden, p. 847, 2018. [Online]. Available: http://www.dymola.com
3. International Energy Agency, "IEA Response System for Oil Supply Emergencies 2012," International Energy Agency, Paris, Tech. Rep., 2012. [Online]. Available: http://www.iea.org/publications/freepublications/publication/EPPD_Brochure_English_2012_02.pdff
4. J. M. Ogden and L. Anderson, "Sustainable Transportation Energy Pathways: A Research Summary for Decision Makers," Institute of Transportation Studies, University of California, Davis, Tech. Rep., 2011. [Online]. Available: http://www.its.ucdavis.edu/?page_id=10063&pub_id=1499
5. European Commission, *A sustainable future for transport - Towards an integrated, technology-led and user-friendly system.* Luxembourg: Publications Office of the European Union, 2009.
6. United Nations Environment Programme, "Hybrid Electric Vehicles: And overview of current technology and its application in developing and transitional countries," United Nations Environment Programme, Nairobi, Tech. Rep., 2009. [Online]. Available: https://www.globalfueleconomy.org/transport/gfei/autotool/approaches/technology/Hybrid%20Electric%20Vehicles%20final%20cs2.pdf
7. G. Pasaoglu, M. Honselaar, and C. Thiel, "Potential vehicle fleet co2 reductions and cost implications for various vehicle technology deployment scenarios in europe," *Energy Policy*, vol. 40, no. C, pp. 404–421, 2012. [Online]. Available: https://EconPapers.repec.org/RePEc:eee:enepol:v:40:y:2012:i:c:p:404-421
8. Fritzson P., *Principles of Object-Oriented Modeling and Simulation with Modelica 2.1.* Wiley-IEEE Press, 2004. [Online]. Available: http://www.wileyeurope.com/WileyCDA/WileyTitle/productCd-0471471631.html
9. M. Association, "Modelica ® - A Unified Object-Oriented Language for Physical Systems Modeling Language Specification," *Interface*, vol. 5, no. 6, p. 250, 2010. [Online]. Available: www.modelica.org

10. Dassault Systèmes AB, Ed., *Dymola - Dynamic Modeling Laboratory - User Manual*. Lund - Sweden: Dassault Systèmes AB, 2012, vol. 1, no. May. [Online]. Available: http://www.dymola.com
11. Modelica Association, "VehicleInterfaces Library (Version 1.2.4) Reference Guide," 2016. [Online]. Available: https://build.openmodelica.org/Documentation/VehicleInterfaces.html
12. O. Tremblay and L. Dessaint, "Experimental validation of a battery dynamic model for EV applications," *World Electr. Veh. J.*, vol. 3, pp. 1–10, 2009.

# Interés de los Usuarios del Sistema Sanitario en relación al uso de las nuevas Tecnologías de la Información y Comunicación en la relación médico-paciente

González Revuelta, María Esther[1]

[1] Tecnologías de la Información y Comunicación. C.H.Torrecárdenas.Almería

**Abstract.** Contextualizamos inicialmente realizando un pequeño resumen del objeto del proyecto para exponer después los avances realizados al respecto en estos últimos meses.
El desarrollo de la Historia clínica Electrónica en el ámbito sanitario es un objetivo sobre el que avanzamos y trabajamos en los últimos años especialmente. La evolución en esta área aporta claros beneficios tanto para el ciudadano como para los profesionales sanitarios.
La disponibilidad de la Historia Clínica del paciente, de sus pruebas, de sus imágenes Radiológicas, de sus datos asistenciales en general independiente de dónde y cuándo sea atendido, a través de Sistemas de Información específicos y de herramientas especialmente diseñadas para ello ha supuesto un gran avance y mejora en la atención sanitaria y está en pleno auge.
Durante este período hemos avanzado principalmente en recoger la información real respecto a la opinión de los pacientes en relación al nivel de interés en conocer el estado de su proceso asistencial a través de las nuevas tecnologías y en algunos casos supliendo la asistencia física a la consulta e in-situ en el hospital o consulta de Atención Primaria. Hemos analizado los datos y en base a la información, actualmente estamos completando con una segunda encuesta más completa que nos permita afinar más en las conclusiones.
.

**Keywords:** mHealth, Aplicaciones móviles, telemedicina

## 1 Introducción

Durante este último periodo hemos avanzado en la recopilación y análisis de la información aportada por los pacientes a través de encuestas tipo likert.
La población objeto de estudio para la primera fase han sido los usuarios del Sistema Sanitario Público de Andalucía que se realizan la prueba del Screening de Cáncer de Cérvix en la consulta de Ginecología de Atención Especializada y mujeres embarazadas que se realizan una extracción de seguimiento.

El Proceso Operativo (clínico-asistencial) se inicia tras el contacto de la persona con el Sistema Sanitario por las diferentes entradas posibles Atención Primaria (AP) o Atención Hospitalaria (AH). La atención que los profesionales ofrecen desde distintos ámbitos de actuación en AP y AH se pone en valor con la continuidad de la asistencia al paciente/familia. Hasta la salida del Proceso Asistencial Integrado (PAI) todo el proceso se desarrolla en el marco establecido por los Procesos Estratégicos y sustentados en los Procesos de Soporte.

La mujer con diagnóstico de cáncer de cérvix uterino no abandona la cadena asistencial, pues precisa de seguimiento periódico. En todo el proceso, pretendemos potenciar el papel del usuario, en este caso, de la mujer que participa y con la cual pretendemos desarrollar una vía de comunicación más directa y cómoda con sus propios resultados.

Las pacientes que tras exploración ginecológica realizada en consultas de AP y/o AH, presentan signos de sospecha clínica y diagnóstica, se les realiza una citología. Si el resultado se considera objeto de estudio se cita en la consulta para realizarle una prueba PCR, en el desarrollo de todo el proceso, desde que se le realiza la prueba hasta que obtiene los resultados transcurre un tiempo, en el cual la mujer desconoce cualquier información positiva o negativa del estado de su prueba. En ocasiones esto provoca un estado de ansiedad, de nerviosismo e impaciencia que podríamos evitar o al menos minimizar en función de los resultados obtenidos.

Más allá de entrar en el proceso de Cáncer de Cérvix que está perfectamente procedimentado, lo que se pretende es minimizar tiempos de espera en la información en situaciones donde los resultados sean negativos, mejorando procesos en cuanto a tiempos, tiempos de espera, tiempos de citación, tiempos de transmisión de resultados... todo ello como soporte a la labor asistencial del profesional y salvaguardando los derechos fundamentales de información, intimidad y confidencialidad.

En el caso del Proceso Asistencial Integrado de embarazo, parto y puerperio, el Proceso Operativo (clínico-asistencial) se inicia tras el contacto de la persona con el Sistema Sanitario por las diferentes entradas posibles (061- AP- AH), y la atención de los profesionales desde diferentes ámbitos de actuación AP y AH. Existe un amplio programa de actividades para hacer seguimiento de forma integrada entre AP y AH, estableciéndose durante todo el proceso una serie de visitas, citas y recomendaciones que fomentan la participación de la mujer en el desarrollo del mismo. Es por tanto, factible que en todo este proceso se haga uso de la tecnología como apoyo y mejora del mismo.
En ambas procesos, seleccionamos uno de los puntos de contacto del usuario con el sistema sanitario, Consulta de Ginecología y Laboratorio.

Para la selección de la muestra se ha utilizado una metodología de muestreo consecutivo, por tanto no probabilístico. El criterio utilizado ha sido el de accesibilidad del paciente a la consulta y aceptabilidad de forma totalmente voluntaria de inclusión en el estudio por parte del paciente.

Los objetivos que nos planteamos son:

- Determinar el grado de interés de los pacientes en conocer información propia de su proceso asistencial a través de las Tecnologías de la Información y Comunicación (TIC) Participando de forma más activa en el desarrollo de su proceso.

- Evaluar el grado de conocimiento de los pacientes inmersos en un Proceso Asistencial de Ginecología y de Embarazo, Parto y Puerperio sobre sus citas próximas para el seguimiento de su proceso mediante los circuitos habituales.

- Acceso de los ciudadanos a la información de su HHCC. Concretamente a la información generada durante un proceso de análisis.

- Ayudar a gestionar procesos, en los que intervienen Atención Primaria y Atención Hospitalaria mediante el uso de las nuevas tecnologías y los medios y herramientas TIC.

- Análisis de procesos intralaboratorio, realizando un seguimiento desde que la muestra entra en laboratorio hasta que se obtiene el resultado, facilitando el conocimiento del estado de la misma por el paciente desde su inicio hasta su fin, permitirá conocer el estatus de la analítica, detectar situaciones de fallo (en caso de que la muestra sea fallida) con antelación, sin necesidad de esperar de forma innecesaria a la próxima visita del usuario.

- Acortar tiempos de espera de los usuarios, centrándonos a priori en aquellos resultados que siendo negativos no son conocidos por el usuario hasta su próxima visita, para la cual podrían transcurrir varios meses. Meses que se convierten en interminables y generan situaciones de ansiedad y preocupación que podríamos evitar.
- En los casos en que los resultados sean positivos,, se podrían agilizar actuaciones, por ejemplo, adelantar la cita, a través de mensajería que se podría enviar de forma automática al profesional implicado.

En relación al análisis estadístico:
Población de estudio: Mujeres acuden a las consultas Ginecología, Cáncer de Cérvix y consulta de laboratorio.
Modo de selección muestral: Selección aleatoria en la consulta diaria durante un periodo de 3 meses hasta alcanzar la muestra.
A partir de una población diana de 540 pacientes. Para conseguir una precisión del 5% en la estimación de una proporción mediante un intervalo de confianza asintótico Normal con corrección para poblaciones finitas al 95%
bilateral, asumiendo que la proporción esperada es del 75% y que el tamaño total de la población es de 540, será necesario incluir 188 pacientes en el estudio.

Las encuestas se realizaron por parte de personas adiestradas en la realización de cuestación. Se realizó inicialmente un pilotaje previo para homogeneizar y consensuar el modo en que se administrarán los cuestionarios y reducir así la variabilidad.

Mediante SPSS se realiza un análisis descriptivo utilizando medidas de tendencia

central y tablas de distribución de frecuencia para las variables cualitativas.

**Estadísticos descriptivos**

| | | Media | Desviación estándar |
|---|---|---|---|
| P1 | Conoce cómo/quién le informará | 3,30 | 1,673 |
| P2 | Conoce tiempo de espera para resultados | 3,58 | 1,537 |
| P3 | Conoce fecha de cita para resultados | 3,86 | 1,677 |
| P4 | Tiene interés en conocer resultados antes de la cita | 4,08 | 1,591 |
| P51 | Preferencia por conocer resultados vía correo electrónico | 2,94 | 2,243 |
| P52 | Preferencia por conocer resultados vía SMS | 1,90 | 2,116 |
| P53 | Preferencia por conocer resultados vía Mensajería Instantánea (Whatsarpp,...) | 2,21 | 2,211 |
| P54 | Preferencia por conocer resultados a través de Página Web | 2,57 | 2,259 |
| P55 | Preferencia por conocer resultados a través de APP | 1,44 | 1,891 |
| P56 | Preferencia por conocer resultados a través de otros medios | ,66 | 1,564 |
| P61 | Interés conocer si la muestra está en laboratorio | 2,93 | 2,042 |
| P62 | Interés conocer si la muestra está procesada | 2,35 | 1,986 |
| P63 | Interés conocer si los resultados están disponibles | 4,37 | 1,408 |
| P64 | Interés conocer si error en muestra | 3,64 | 2,106 |
| P7 | Anticipar el conocimiento de estado de muestra y de resultados alivia ansiedad | 4,08 | 1,258 |
| P8 | Si los resultados son normales evitaría cita médica | 3,78 | 1,656 |
| P9 | Si los resultados son normales y acompaña informe evitaría cita médica | 4,16 | 1,396 |
| P10 | Interés en recibir información hábitos vida saludable relacionados con su caso | 4,61 | ,938 |

Estamos en proceso de análisis de los datos y recopilación de bibliografía existente en relación en definitiva a la opinión de los ciudadanos sobre el uso y aplicación de las TIC en el ámbito Sanitario.

Durante el proceso de análisis decidimos ampliar la variables utilizadas para realizar un estudio más completo y mejorar el foco de estudio.

Generamos una nueva encuesta con las siguientes variables:

Datos de filiación: Sexo, edad, localidad de residencia, código postal

Datos relacionados con su situación profesional: Nivel de estudios, situación profesional.

Datos de estado de salud personal y hábitos

Entorno familiar

Nivel de uso de Internet y nuevas tecnologías

Interés y confianza en el uso de medios tecnológicos digitales para el seguimiento de su proceso clínico.

En este caso para el tamaño y procedimiento de muestra: Para conseguir una precisión del 5% en la estimación de una proporción mediante un intervalo de confianza asintótico normal al 95% bilateral, asumiendo que la proporción de interés para conocer los resultados antes de la cita del médico es del 60% será necesario incluir 369 individuos en el estudio.

Se realizará un muestreo estratificado con afijación proporcional.

Estamos en fase de recogida de la información para proceder posteriormente a su análisis e incorporación al desarrollo del proyecto.

Como actividades externas mencionar la asistencia al Congreso de la Sociedad Andaluza de Calidad Asistencial, realizado en el mes de Noviembre. Este congreso está estrechamente relacionado con el tema de análisis y desarrollo.
La principal área temática de este congreso fue la reflexión sobre las Redes de Innovación en Salud, teniendo como eje los planteamientos sobre los que se trabaja en relación al trabajo colaborativo en las redes de conocimiento, redes asistenciales, redes de investigación, redes virtuales y redes de resultados.
He participado además de asistente, como miembro del Comité Científico y como moderadora de una mesa de comunicaciones titulada "Humanización y Participación".
Añadir la asistencia a varias Jornadas también de interés:

-Jornadas Andaluzas de Salud Investiga en Octubre organizado por la Consejería deSalud y Fundación Progreso y Salud.

- Jornada Ciencia de Datos y Big Data en Salud, organizado por la Escuela Andaluza de Salud Pública.

# Distribución del rate para el codec Motion Compensated JPEG2000

J.C. Maturana-Espinosa*

*University of Almería Ctra. Sacramento, s/n Almería, 04120, Spain

**Resumen** MCJ2K (Motion Compensated JPEG2000) es un códec de vídeo basado en MCTF (Motion Compensated Temporal Filtering) y J2K (JPEG2000). MCTF analiza una secuencia de imágenes, generando una colección de sub-bandas temporales, las cuales están comprimidas con J2K. Comparado con el MJ2K (Motion JPEG2000), el rendimiento D/R (Distorsión/Rate) del MCJ2K es mejor que el del MJ2K, porque la redundancia temporal mostrada por la mayoría de las secuencias puede ser eliminada por el MCTF. Además, las corrientes de código MCJ2K pueden ser servidas por servidores JPIP (J2K Interactive Protocol) estándar, gracias al uso exclusivo de formatos de archivo estándar J2K. En escenarios con ancho de banda limitado, un problema importante en el MCJ2K es determinar la cantidad de datos de cada sub-banda temporal (el número de sub-capas de cada sub-banda) que deben ser transmitidas, con el fin de maximizar la calidad de las reconstrucciones en el lado del cliente. Para resolver este problema, hemos propuesto dos algoritmos de distribución del rate, que proporcionan reconstrucciones de calidad progresivas. El primero, FSO (Full Search Optimization), determina la mejor progresión de las capas para cada sub-banda, en términos de la siguiente capa para añadir, pero es computacionalmente costoso. El segundo, PTL (Progressive Transmission by subband-Layers) es subóptimo, pero mucho más rápido, y es más conveniente para escenarios de streaming en tiempo real. Una comparación experimental muestra que, incluso utilizando un esquema de compensación de movimiento directo, el rendimiento D/R del MCJ2K es competitivo no sólo en comparación con el MJ2K, sino también con otros códecs de vídeo.

**Keywords:** MCJ2K, MCTF, rate

## 1. Introducción

MCJ2K[1] (Motion Compensated JPEG2000) es una combinación de dos etapas fundamentales: (1) MCTF[2] (Motion Compensated Temporal Filtering) y (2) J2K. Básicamente, MCTF es una transformación que introduce una secuencia de imágenes y emite una secuencia de *MCTF-coeficientes* (*coefs* en el resto del documento), agrupados en una colección de subbandas temporales. Luego, estos coefs son comprimidos con J2K, resultando en una colección de flujos de código J2K que pueden ser transmitidos usando JPIP. El rendimiento D/R (Distorsión/Tasa) del MCJ2K puede ser claramente mejor que el del J2K, dependiendo

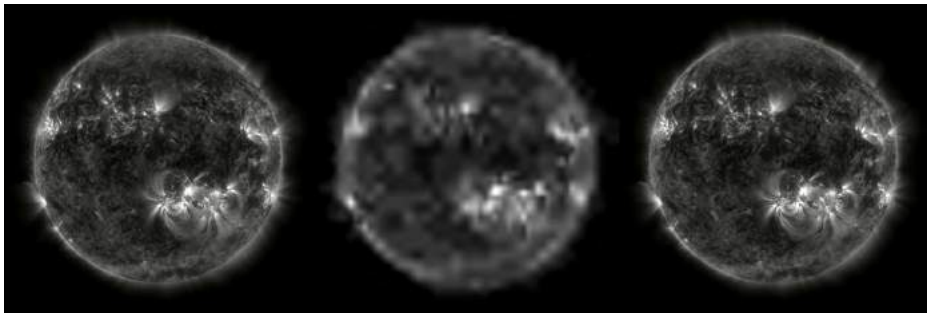


Figura 1: Izquierda: imagen original del Sol con $512 \times 512$ y una cadencia de 1/12 imágenes/segundo. Centro: misma imagen descomprimida con J2K a 0,08 kbps. Derecha: misma imagen descomprimida con MCJ2K a 0,04 kbps. Crédito de la imagen: NASA/SDO/AIA.

de la correlación temporal entre las imágenes de entrada. A modo de ejemplo, la figura 1 muestra una imagen del Sol (de una secuencia) descomprimida con MJ2K y MCJ2K, a velocidades de bits similares.

El MCJ2K es una extensión directa del J2K, y ya ha sido propuesto anteriormente. Sin embargo, la adaptación del MCJ2K a los servicios JPIP estándar, como el Helioviewer, es un trabajo novedoso. Además, dos nuevos RA (Rate-Allocation[1]) algoritmos: Se han propuesto FSO (Full Search Optimization) y PTL (Progressive Transmission by subband-Layers) y se han evaluado experimentalmente. Ambos algoritmos se ejecutan en el momento de la poscompresión para determinar la buena (óptima, si es posible) progresión de las capas sub-banda (las capas de calidad con el mismo índice de cada codo de una sub-banda temporal conforman una capa sub-banda). Así, cuando un cliente JPIP solicita una WOI de una secuencia de imágenes a un servidor JPIP, el servidor envía (utilizando estrictamente la funcionalidad JPIP) la información necesaria para que el cliente pueda determinar la progresión de dichas capas de sub-banda.

Nuestra implementación de MCJ2K[2] es un bucle-abierto de estructura t+2D (ver Fig. 2a) ), donde el filtro temporal (un 1/3 lineal 1D-DWT) se aplica primero, y una 2D-DWT (dada por el estándar del codec J2K) se aplica a las imágenes transformadas (coefs). La etapa "t" corresponde a un proceso de $T$-niveles MCTF (denotado en la figura por MCTF$^T$) que explota la redundancia temporal de la secuencia de imágenes, y la etapa"2D" corresponde a un banco de compresores MJ2K, que explotan la redundancia espacial en cada coef, y realizan codificación de la entropía. En la figura, $s$ representa la secuencia original

[1] El término "rate-allocation" se refiere a la acción de clasificar el código de flujo con el fin de proporcionar algún tipo de escalabilidad. El término "rate-control" se usa cuando el codificador decide qué información es representada por el flujo de código en escenarios con restricciones de velocidad.

[2] Disponible en https://github.com/vicente-gonzalez-ruiz/MCTF-video-coding

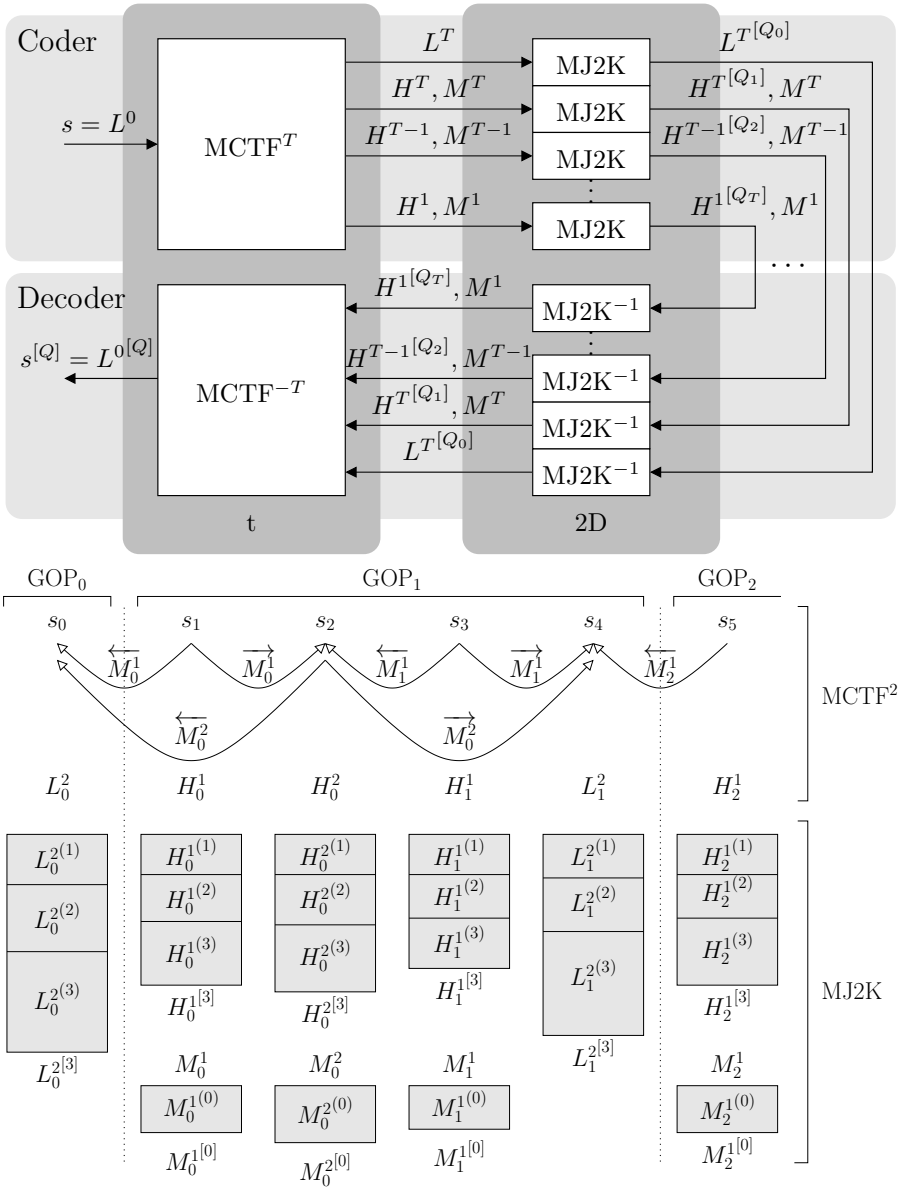



Figura 2: Etapas MCJ2K: a) MCTF b) A collection of compressed textures.

y $[s]^Q$ una aproximación progresiva de $s$, reconstruida usando capas de calidad $Q$. MCTF$^T$ representa $s$ como una colección de $T+1$ texturas-subbandas temporales $\{L^T, \{H^t; 1 \leq t \leq T\}\}$ y $T$ vectores de movimiento-"subbandas" $\{M^t; 1 \leq t \leq T\}$.

Actualmente se ha optimizado el código, haciendo que cada imagen se trate fisicamente de forma independiente, aumentando así el rendimiento en la codificación.

Fig. 2b) muestra un ejemplo de la organización de un flujo de código MCJ2K. Nueve imágenes han sido comprimidas (aunque sólo las primeras seis $s_0, \cdots s_5$ se muestran), el tamaño del GOP es $G = 4$ (es decir, $T = 2$ ($G = 2^T$)), excepto para el primer GOP que siempre tiene una sola imagen. MCTF$^2$ transforma la secuencia de entrada $s$ en sub-bandas de textura de tres niveles de resolución y de 2 de movimiento, resultando: $\{L^2, H^2, M^2, H^1, M^1\}$. $L^2$ es la sub-banda de textura de baja frecuencia, y representa los componentes temporales de baja frecuencia de $s$. $\{H^2, H^1\}$ son las sub-bandas de textura de alta frecuencia que contienen las componentes temporales de alta frecuencia de $s$. $\{M^2, M^1\}$ son los vectores de movimiento. En Fig. **??**, las flechas indican la dirección del proceso de EM. Cuando se aplica la transformación inversa, se genera una sucesión de incrementos temporales $\{L^2, L^1, L^0\}$ de la secuencia original.

El resto de este documento se organiza del siguiente modo: la Sección 2 muestra nuestra propuesta e introduce técnicas de evaluación del algoritmo propuesto que están ahora ultimándose. Las conclusiones principales de aplicar nuestra propuesta se muestran en la Sección 3.

## 2. Avances

### 2.1. Distribución del rate Post-compresión

En el momento de la descompresión, el orden en el que se encuentran las subcapas de la subbanda recuperada desde el servidor JPIP debería minimizar la curva D/R, para cualquier velocidad de bits. Para esta tarea, proponemos los dos enfoques siguientes.

**Búsqueda completa (FSO)** La contribución de una subcapa temporal de movimiento a la calidad de la reconstrucción depende en gran medida del rendimiento del proceso de EM, que, especialmente en aquellos casos en los que el movimiento real de la escena no ha sido determinado con precisión, puede mostrar un comportamiento no lineal[3]. Esto se muestra en los resultados cuando, después de enviar una subbanda de movimiento, la calidad de la reconstrucción empeora. Por esta razón, después de recuperar, por ejemplo, la capa de calidad

[3] lo que significa que la energía de las subbandas $\{H^t; 1 \leq t \leq T\}$ no es proporcional a la diferencia entre las subbandas de movimiento calculadas y el movimiento real de la secuencia. En otras palabras, que un pequeño error en los vectores de movimiento puede producir un error de predicción mayor que un error en los vectores de movimiento

$L^{T\,1}$ (que siempre contribuye a la calidad de la reconstrucción más que cualquier otra subcapa, se prueban varias alternativas $\{M^T, H^{T\,1}, H^{T-1\,1}, \cdots, H^{1\,1}\}$ para determinar la siguiente subcapa de cualquier subcadena con la mayor contribución posible. Si ocurre que $\lambda_{M^T} > \lambda_{H^{t\,1}}, \forall t = T, \cdots, 1$, la siguiente subcapa para decodificar debería ser $M^T$ y el siguiente grupo de alternativas sería $\{M^{T-1}, H^{T\,1}, H^{T-1\,1}, \cdots, H^{1\,1}\}$. De lo contrario, si por ejemplo, $\lambda_{H^T} > \lambda_{M^t}$ y $\lambda_{H^T} > \lambda_{H^t}, \forall t = T-1, \cdots, 1$, después $L^{T\,1}$ la siguiente capa decodificada debe ser $H^{T\,1}$, y el siguiente conjunto de alternativas sería $\{M^T, H^{T\,2}, H^{T-1\,1}, \cdots, H^{1\,1}\}$.

FSO es un algoritmo codicioso que implementa esta idea, y determina, GOP por GOP, el orden óptimo de las capas entre todas las sub-bandas. Para ello, se evalúan todas las alternativas posibles, suponiendo que un flujo de codestream MCJ2K puede truncarse en cualquier subcapa, y que sólo es posible una orden de transmisión. Esta última premisa es importante, porque, si la velocidad de transmisión se conoce a priori, se puede encontrar un orden con una mejor pendiente D/R que la encontrada por el algoritmo FSO.[4] Una descripción del pseudocódigo de FSO sería:

1. for each GOP:
2. `retrieve` $L^{T\,1}$
3. $S = \{M^T, M^{T-1}, \cdots, M^1\}$
4. for each $q \in \{1, \cdots, Q\}$:
5. $S = S \cup \left\{L^{T\,q+1}, \{H^{Q\,q}, H^{T-1\,q}, \cdots, H^{1\,q}\}\right\}$
6. while $S \neq \emptyset$:
7. `find` $S_i \in S$ s.t. $\lambda_{S_i} \geq \lambda_{S_j}\ \forall S_j \in S \setminus \{S_i\}$
8. `send` $S_i$
9. $S = S \setminus \{S_i\}$

La operación `send` escribe en una lista el orden en el que deben transmitirse las subcapas y esta lista se envía en un segmento COM del encabezado J2K del coef correspondiente de $L^T$ (un coeficiente por cada GOP). FSO ejecuta $\mathcal{O}((QT)^2)$ veces la operación `find`, y cada hallazgo implica una nueva reconstrucción del GOP. Por lo tanto, FSO es muy exigente y difícil de ejecutar en tiempo real.

**Transmisión Progresiva por capas de subbandas (PTL)** Pseudo-código de PTL:

[4] Considere que la calidad de la reconstrucción no siempre es proporcional a la cantidad de datos decodificados.

1. for each GOP:
2. `read` $\Lambda = \{ \lambda_{L^T} = \{\lambda_{L^T 1}, \cdots, \lambda_{L^T Q}\}, \lambda_{H^T} = \{\lambda_{H^T 1}, \cdots, \lambda_{H^T Q}\}, \vdots \lambda_{H^1} = \{\lambda_{H^1 1}, \cdots, \lambda_{H^1 Q}\} \}$
3. for each $t \in \{T, \cdots, 1\}$:
4. $\lambda_{H^t} = \lambda_{H^t} - \lambda_{H^t Q}$
5. $\lambda_{H^t} = \lambda_{H^t} \times \beta_{H^t}$
6. `sort`, in descending order, $\Lambda$ by slopes
7. for each $t \in \{T, \cdots, 1\}$:
8. `insert` $M^t$ in $\Lambda$
9. `send` $\Lambda$ by slopes

PTL puede funcionar en ambos extremos de un sistema JPIP. Si PTL se ejecuta en el lado del servidor, la operación `send` debería escribir en el segmento COM de cada coef de $L^T$ la lista de sub-bandas tal y como se encuentran en la versión ordenada de $\Lambda$ (línea 6 del algorithm). Alternativamente, en el segmento COM se pueden almacenar los valores originales $\Lambda$ (línea 2 del algoritmo), y el algoritmo se puede ejecutar en el lado del cliente. En cualquier caso, el tiempo de ejecución de PTL es insignificante y, por lo tanto, puede utilizarse en escenarios de streaming en tiempo real, donde se generan las secuencias de imágenes y, a continuación, se transmiten con un retraso mínimo.

### 2.2. Rendimiento de los algoritmos propuestos

Los algoritmos han sido evaluados en términos de RD, donde el parámetro $Q$ controla la precisión de los algoritmos RA.

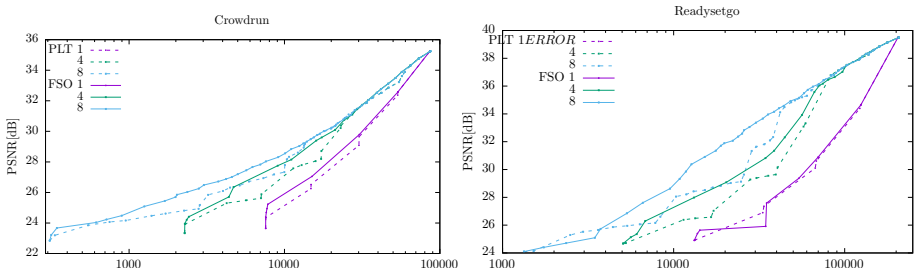



Figura 3: Rendimiento de FSO a diferente número de capas de calidad $Q = \{1, 4, 8\}$ ofrecen un buen rendimiento RD.

En la Fig. 3, a más capas de calidad mayor rango de bit-rate con el que puede enviarse el video, pudiendo enviarlo a un bit-rate más bajo, ya que se dispone de capas más pequeñas. Aunque PLT demuestra un muy buen desempeño si no sólo se evalúa el inicio de la lista ordenada de capas, sino en su totalidad.

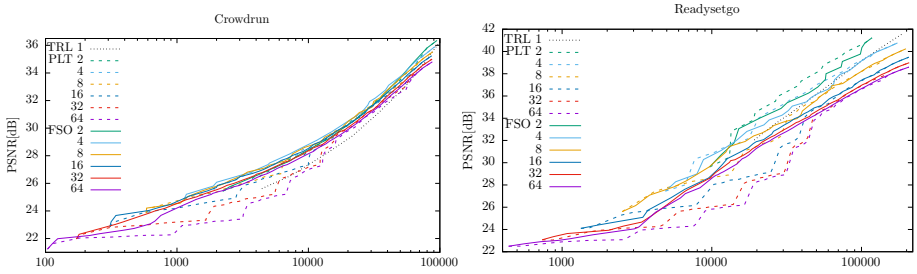



Figura 4: Rendimiento de FSO para diferente número de TRLs. $G = \{2, 4, 8, 16, 32, 64\}$.

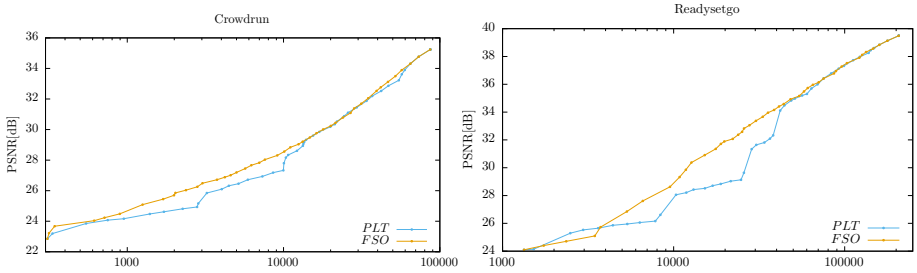



Figura 5: PLT ofrece un buen rendimiento frente a FSO.

Fig. 4 La división del codestream en un mayor o menor número de TRLs repercute en el rango de bit-rate y PSNR.

A menor número de TRLs el bit-rate necesario para enviar una capa del codestream es mayor (y significativo en los resultados). Aunque, al enviar todo el codestream se alcanza mayor PSNR (poco notorio). Y viceversa.

Un valor que suele resultar adecuado es 5 TRLs. Aunque esto depende de la cantidad y la predicibilidad del movimiento que presente el vídeo.

## 3. Conclusiones

MCJ2K es un códec eficiente y totalmente compatible con JPIP para la transmisión interactiva de vídeo, y ofrece una alternativa al MJ2K si se puede sacrificar algún grado de escalabilidad temporal. La relación de compresión de MCJ2K es similar a otros estándares de codificación de vídeo escalables como SHVC.

El rendimiento de la DR en el MCJ2K puede mejorarse significativamente en tres aspectos principales: (1) un esquema ME/MC más preciso, (2) el uso de escalabilidad espacial/calidad al codificar los datos de movimiento y el desarrollo de nuevos algoritmos RA, y (3) el uso de esquemas de codificación en los que la información de movimiento puede estimarse en el decodificador (para evitar

enviarla como parte del flujo de código). Esto se puede llevar a cabo en aquellos contextos en los que el movimiento a gran escala es predecible, tales como secuencias de imágenes del Sol, cuya velocidad de rotación es estable y buena conocido.

## Referencias

1. T. Tuithung, S.K. Ghosh, and Jayanta Mukherjee. Motion compensated JPEG2000 based video compression algorithms. *International Journal of Signal and Imaging Systems Engineering*, 1(3/4):197 – 212, 2008.
2. R. Xiong, J. Xu, F. Wu, and S. Li. Adaptive MCTF based on correlation noise model for SNR scalable video coding. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1865–1868, July 2006.

# Atravesando NAT Simétricos en redes P2P mediante predicción de puertos colaborativa

Cristóbal Medina-López

Universidad de Almería, Sacramento S/N, Almería, España,
cristobalmedina@ual.es

**Resumen** Los NATs (Network Address Translators) se usan para incrementar el número máximo de hosts conectados en IPv4, y para mejorar la seguridad usándolos como cortafuegos. Sin embargo, los NATs complican las comunicaciones entre los procesos que se están ejecutando tras ellos. Para tratar con este problema, se han desarrollado algunos protocolos como STUN (Session Traversal Utilities for NAT) y ICE (Interactive Connectivity Establishment), que hacen posible atravesar NAT (no simétricos) mediante descubrimiento de endpoints públicos y técnicas de UDP Hole Punching (UHP). Desafortunadamente, en los NAT simétricos, los procesos privados usan diferentes puertos públicos para cada combinación de endpoints y por tanto, UHP necesita técnicas adicionales como la predicción de puertos para conseguir atravesarlos. En este trabajo se presenta un algoritmo denominado SNT-CPP (Symmetric NAT Traversal using Collaborative Port Prediction) que amplía las funcionalidades de STUN y ICE permitiendo atravesar algunos NAT simétricos.

**Keywords:** Symetric NAT, P2P, UHP, predicción de puertos.

## 1. Introducción

En los sistemas P2P (Peer-to-Peer) reales, la mayoría de los peers se ejecutan detrás de dispositivos NAT [11]. Un NAT es un tipo de router que interconecta dos redes diferentes, normalmente una privada (o interna) y otra pública (o externa). Mediante una colección de entradas de tabla, estos dispositivos establecen una correspondencia entre los pares IP:Puerto de la red pública y privada.

Los dispositivos NAT se usan principalmente por dos razones: (1) para reducir el número de direcciones IP públicas usadas por los hosts privados, y (2) para incrementar la seguridad de la red privada, gracias a que el tráfico originado desde el lado público del NAT será bloqueado a menos que una entrada de correspondencia exista en la tabla del NAT. Estas entradas pueden ser creadas de forma estática por el administrador del NAT, o de forma dinámica teniendo en cuenta el tráfico saliente. Los NATs se pueden clasificar en diferentes tipos dependiendo de cómo administren la asignación en las entradas de la tabla [6,9]:

- NATs de tipo cono (CN). Todo el tráfico generado por el endpoint interno $X$ es mapeado por un algoritmo de mapeo de puertos (PMA) a un único puerto

fuente externo $s(X)$ del NAT independientemente del endpoint externo de destino $Y$.
- NATs de tipo simétrico (SN). El PMA usa un puerto externo diferente $s(X, Y)$ para cada combinación $(X, Y)$. Por tanto, una combinación de IP:Puerto externa puede ser usada para comunicarse con un único endpoint fuera del NAT.

Además, dependiendo del algoritmo de filtrado del NAT, un NAT de tipo cono puede ser categorizado en [12]:

- *Full Cone NATs (FCN)*, o NATs de tipo cono completo. Son aquellos en los que cualquier paquete recibido desde la red pública del NAT a través del puerto origen $s$ es reenviado al endpoint interno $X$ sin importar el endpoint público $Y$. Es decir, no se usa ningún filtrado.
- *Restricted Cone NATs (RCN)*, o NATs de tipo cono restringido. Son aquellos en los que los paquetes recibidos por el NAT a través de $s$ son reenviados a $X$ solo si ha existido tráfico previamente de $X$ a $Y.a$ (dirección de $Y$). Este comportamiento es denominado "filtrado por dirección".
- *Port-Restricted Cone NATs (PRCN)*, o NATs de tipo cono restringido por puerto. Son RCNs que también comprueban el puerto origen del tráfico: un paquete recibido a través de $s$ desde $Y$ es reenviado a $X$ solo si ha existido tráfico previo desde $X$ hacia $Y$. Este comportamiento es denominado "filtrado por puerto".

Finalmente, dependiendo del algoritmo de mapeo de puertos usado, los NATs pueden ser también clasificados en los siguiente tipos [12]:

- NATs con *Port-Preservation Allocation (PPA)* [3] o asignación por conservación de puertos. Mediante este algoritmo se intenta que el puerto origen sea el mismo que el puerto del endpoint interno $s = X.p$. Estos NATs también son conocidos como NAT de asignación estática de direcciones [10]. En el caso de que $X.p$ esté ya en uso, se asignará el siguiente puerto libre. En este caso, nos referimos a NAT con asignación secuencial de puertos o *Sequential-Port Allocation (SPA)* [3] .
- NATs con *Random-Port Allocation (RPA)* o asignación aleatoria de puertos. En este algoritmo se asigna $s$ de forma aleatoria de entre los puertos libres del NAT [4].

En el caso del tráfico UDP, la mayoría de las técnicas para atrevar NATs se basan en *UDP Hole Punching (UHP)* [6]. UHP se basa en la idea de que, para permitir la recepción de un paquete entrante desde un peer externo (usando su endpoint) $Y$, un peer interno ("NATed") $X$ puede crear una entrada en la tabla del NAT asociada al puerto origen $s$ enviando al menos un paquete UDP hacia $Y$.

Aunque se han propuesto técnicas completamente autónomas para UHP [7], la mayoría de las soluciones (como las basadas en STUN [8]) usan al menos un

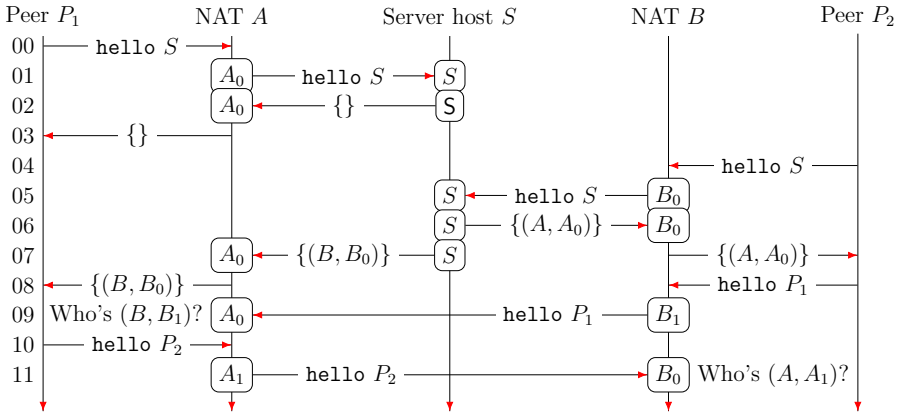



Figura 1: Un ejemplo que muestra por qué no es posible establecer una conexión cuando dos peers $P_1$ y $P_2$ está detrás de NATs simétricos.

servidor público para detectar la presencia de NATs, e introducir los NATed peers que llevan a cabo el UHP.

UHP funciona bien con CNs, pero desafortunadamente falla cuando ambos NATs son SN [2] o una combinación de PRCN y SN [12]. La Figura 1 muestra un ejemplo de este fallo. Como se observa, la comunicación falla porque en el paso 11 el NAT $\mathcal{B}$ solo reenvía el tráfico entrante que proviene del servidor $S$ (una situación similar ocurre en el NAT $\mathcal{A}$). Por tanto, el problema a resolver es predecir el resultado de PMA en el NAT del peer al cual nos queremos conectar.

| Peer1/2 | FCN | PRCN | SN-SPA | SN-RPA |
|---|---|---|---|---|
| FCN | si | si | si | si |
| PRCN | si | si | *(si)* | no |
| SN-SPA | si | *(si)* | *(si)* | no |
| SN-RPA | si | no | no | no |

Cuadro 1: Éxito teórico de atravesar NATs con diferentes configuraciones.

El dinamismo en la creación de endpoints en la red interna de los peers hacen la predicción difícil para algunas combinaciones de NATs. La Tabla 1 resume el éxito esperado para las técnicas basadas en UHP. La entradas etiquetadas como "si" hacen referencia a usar UHP sin predicción de puertos. Las etiquetadas con "no" indican aquellas combinaciones que, por lo general, sin usar predicción de puertos no permitirán la comunicación. Las marcadas como *(si)* son combinaciones relevantes para el desarrollo de este paper.

El resto de este documento se organiza del siguiente modo: la Sección 2 muestra nuestra propuesta e introduce técnicas de evaluación del algoritmo propuesto

que se llevarán a cabo en estudios posteriores. Las conclusiones principales de aplicar nuestra propuesta se muestran en la Sección 3.

## 2. Avances

Nuestra propuesta, SNT-CPP, se basa en la idea de que, usando un alto número de servidores públicos, la predicción de puertos puede ser más exacta. SNT-CPP está diseñado para ser usado en sistemas P2P, como son sistema de video conferencia de navegador a navegador [1] o el protocolo P2PSP [5]. En el P2PSP, un peer entrante, después de contactar con el *splitter* transmitirá mensajes de *[hello]* a al menos $m \leq T$ peers monitores, los cuales pueden actuar como servidores públicos reportando información sobre el NAT del peer entrante al resto del team. Por simplicidad, el resto de esta sección se enmarcará en el protocolo P2PSP.

### 2.1. UHP Handshake

SNT-CPP define un conjunto de pasos llevados a cabo por todos los miembros del team, estos son ejecutados cuando un nuevo peer $P_i$ se une al equipo. Primero, $P_i$ debe contactar con el splitter $S$. Tras ello, $S$ envía a $P_i$ la lista de peers actual, y añade a $P_i$ al final de su lista. Con el objetivo de crear una entrada en la tabla NAT de $P_i$ para el tráfico UDP entrante, $P_i$ envía mensajes de [`hello`] a $S$ y a todos los peers del team, los cuales responden al peer que se acaba de incorporar. Cada uno de estos mensajes es respondido con un mensaje [`ack`]. En caso de superar un número de fallos $P_i$ no sería incorporado al team, si por el contrario todo va bien, $P_i$ comunica a $S$ que su incorporación fué exitosa. Mientra tanto, los peers previamente añadidos reciben información sobre el NAT desde $S$ que es generada por $P_i$ quién realiza la predicción de puertos y envía por cada uno de los puertos predichos un mensaje [`hello`]. Los monitores se comportan como peers normales, excepto porque son contactados por $P_i$ primero, y por tanto, deben responder a este evento.

La Figura 2 muestra un ejemplo detallado de como funciona el UHP usado en SNT-CPP en condiciones ideales cuando el primer puerto predicho en ambos extremos es correcto. El nodo $S$ (ejecutándose en el host con dirección IP $\mathcal{S}$) usa el puerto público $\mathsf{S}$, y el peer monitor $M$ (que se ejecuta en el host con dirección IP $\mathcal{M}$), el puerto público $\mathsf{M}$. Cuando llegan dos nuevos peers $P_1$ y $P_2$ que están tras sendos NATs ocurre lo siguiente:

00. $M$ solicita unirse al team, y $S$ envía a $M$ una lista de peers vacía. En este momento, $M$ se ha unido al team.
01. $P_1$ solicita a $S$ unirse a través de un puerto externo $\mathsf{A0}$. $S$ envía a $P_1$ la lista de peers. Esta lista contiene únicamente el endpoint $(\mathcal{M}, \mathsf{M})$.
02. NAT $\mathcal{A}$ reenvia hacia $P_1$ el mensaje anterior.
03. $P_1$ responde con [`hello` $(\mathcal{M}, \mathsf{M})$] a $M$.
04. $\mathcal{A}$ reenvía el mensaje anterior, que es recibido por $M$. Como $\mathcal{A}$ es un NAT simétrico, se usa un nuevo puerto fuente $\mathsf{A1}$ para ese mensaje.

05. $M$ envía [`ack` $(\mathcal{M}, \mathsf{M})$] hacia $(\mathcal{A}, \mathsf{A1})$.
06. El mensaje anterior es reenviado por $\mathcal{A}$. Al mismo tiempo, $M$ informa a $S$ de que $P_1$ se ha comunicado con el, usando el puerto externo $(\mathcal{A}, \mathsf{A1})$.
07. $S$ confirma la recepción del mensaje anterior.
08. $P_2$ solicita unirse al team y $S$ le envía la lista de peer actual, que contiene el endpoint de $M = (\mathcal{M}, \mathsf{M})$ y la tupla $((\mathcal{A}, \mathsf{A0}), \Delta_{\mathcal{A}}, \#P_2)$ (el endpoint externo usado por $P_1$ para comunicarse con $S$, la *la distancia máxima al siguiente puerto* en el NAT $\mathcal{A}$, $\Delta_{\mathcal{A}}$ medida por $S$ para $P_1$ durante su incorporación en el team, y el índice de $P_2$, $\#P_2$, en la lista de peers). Usando esta información, $P_2$ llevará a cabo la predicción para el puerto externo que $\mathcal{A}$ debería asignar a $P_1$ cuando se comunique con $P_2$. Esta predicción es una lista de puertos $Z$.
09. $\mathcal{B}$ retransmite el mensaje anterior.
10. $P_2$ envía un mensaje [`hello` $M$] a $\mathcal{M}$.
11. $\mathcal{B}$ retransmite el mensaje anterior, el cual llega a $M$, y $P_2$ envía [`hello` $(\mathcal{A}, \mathsf{A2})$] hacia ($\mathcal{A}$,A2), que ha sido calculado en el paso 08.
12. El mensaje anterior llega a $(\mathcal{A}, \mathsf{A2})$ (lo cual es correcto), pero $\mathcal{A}$ descarta este paquete porque aún no hay una entrada en la tabla del NAT para la clave $((\mathcal{B}, \mathsf{B2}), \mathsf{A2})$.
13. $M$ confirma el mensaje [`hello` $M$] que llega en el paso 11.
14. El mensaje [`ack` $M$] es recibido por $P_2$ y $M$ informa a $S$ que $P_2$ está también usando el puerto B1. Esta información es utilizada para calcular la distancia máxima de salto entre puertos $\Delta_{\mathcal{B}}$ en el NAT $\mathcal{B}$, medida por $P_2$ durante su incorporación.
15. $S$ confirma la recepción del mensaje anterior.
16. $S$ envía a $P_1$ el mensaje $[(\mathcal{B}, \mathsf{B0}), \Delta_{\mathcal{B}}, \mathsf{S}']$ (endpoint externo usado por $P_2$ para comunicarse con $S$, medida de salto para $P_2$ y un nuevo puerto de escucha temporal $\mathsf{S}'$ en $S$). La tupla $((\mathcal{B}, \mathsf{B0}), \Delta_{\mathcal{B}})$ permite a $P_1$ predecir que puerto externo (B2) debería usarse en NAT $\mathcal{B}$ cuando $P_2$ envía un paquete a $P_1$. El socket extra de $S$ a $\mathsf{S}'$ será usado para actualizar el puerto externo que $P_1$ está usando actualmente para comunicarse con el resto de peers del team.
17. $P_1$ recibe el mensaje anterior.
18. $P_1$ envía un mensaje [`hello` $P_2$] a EEP (NAT $\mathcal{B}, \mathsf{B2}$).
19. $P_1$ envía un mensaje [`hello` $S$] a EEP $(\mathcal{S}, \mathsf{S}')$.
20. NAT $\mathcal{B}$ reenvía el mensaje [`hello` $P_2$] a $P_2$ y [`hello` $S$] es recibida por $S$ (quien actualiza el puerto externo para $P_1$). Hay que tener en cuenta que en este momento, $P_2$ sabe que $P_1$ puede comunicarse con el.
21. Ambos, $S$ y $P_2$ confirman los mensajes [`hello`].
22. [`ack` $S$] es recibido por $P_1$, [`ack` $P_2$] es recibido por NAT $\mathcal{A}$ y como el timer asignado al mensaje [`hello` $P_1$] enviado en el paso 11 se agota, este mensaje es reenviado.
23. $P_1$ recibe [`ack` $P_2$] y NAT $\mathcal{A}$ recibe [`hello` $P_1$].
24. [`hello` $P_1$] es enviado a $P_1$. En este momento, $P_1$ sabe que $P_2$ puede comunicarse con el.
25. $P_1$ confirma la recepción del mensaje anterior [`hello` $P_1$].
26. [`ack` $\mathtt{P}_1$] llega a NAT $\mathcal{B}$.

27. [ack $P_1$] llega a $P_2$.
28. $P_1$ y $P_2$ comunican a $S$ el puerto fuente usado por el otro peer.
29. Esta información es recibida por $S$, que actualiza la información del puerto externo para $P_1$ y $P_2$.

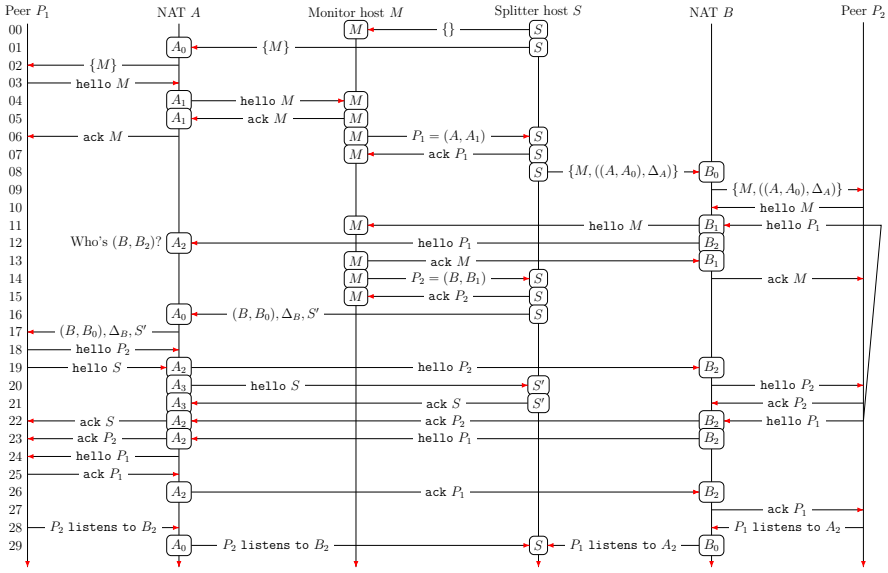



Figura 2: Línea de tiempo de una interacción UHP ideal entre dos peers $P_1$ y $P_2$ que están tras NAT simétricos.

### 2.2. Algoritmo de predicción de puertos

Por cada peer $P_j$ que se incorpora se lleva a cabo una predicción de puertos después de recibir un mensaje de $S$ indicando que $P_i$ quiere unirse al team (ver paso 16 en la Figura 2). Este mensaje contiene (para el NAT $\mathcal{B}$):

1. El puerto externo $(\mathcal{B}, \mathtt{B0})$ que $P_i$ usa para comunicarse con $S$.
2. La distancia máxima de salto entre puertos $\Delta_{\mathcal{B}}$, que determinará el resto de puertos probados por el peer $P_j$.

La salida de este algoritmo es una lista ordenada de puertos $Z_i$ (que es diferente para cada $P_j$).

La predicción de puertos se ha diseñado considerando que:

1. $S$, así como todos los peers del team (incluyendo $P_i$) tienen exactamente la misma lista de peers, sin incluirse así mismos.

2. El $\Delta$ medido puede ser diferente de 1 por dos motivos: (1) porque el NAT no asigna los puertos secuencialmente, y (2) porque el NAT sí que asigna los puertos secuencialmente, pero, durante la medida del $\Delta$, había otros procesos corriendo tras el NAT que usaron algunos puertos. La consideración de un $\Delta >= 1$ mejorará la eficiencia en la predicción minimizando el número de pasos UHP hasta conseguir la incorporación de $P_i$.

Bajo estas consideraciones, la lista de puertos predichos que un peer $P_x$ lleva a cabo se determina por:

$$\begin{array}{rcl} Z_x & = & \mathsf{A}_0 + x + \{s \in \{0, 1, \cdots, N/2 - 1\}\}; \\ Z_x & += & \mathsf{A}_0 + (x + \{s \in \{0, 1, \cdots, N - 1\}\}) \cdot \Delta. \end{array} \quad (1)$$

donde "+ =" denota la concatenación de la lista, $N$ es el número de puertos acertados, $\mathsf{A}_0$ es el primer puerto externo (el usado para comunicarse con $S$) asignado al peer entrante y $\Delta$ es la distancia (máxima) de salto entre puertos establecida en el NAT del peer entrante.

## 3. Conclusiones

La predicción de puertos colaborativa puede estimar de forma precisa la distancia para el siguiente salto del puerto asignado en NATs de tipo SN-SP, por lo que se necesita un número bajo de puertos adivinados para tener éxito atravesando el NAT. Por otro lado, un alto número de peer colaborando permiten obtener una mayor tasa de éxito en comparación con otras soluciones que usan servidores públicos, especialmente para distancias de salto mayores a uno. Nuestra propuesta podría aplicarse a soluciones peer-to-peer estandar para la Web (por ejemplo WebRTC), como una última oportunidad antes de usar servidores TURN para tratar con NAT simétricos, lo cuál es actualmente la única alternativa. Como trabajo futuro, sería interesante una implementación para estos entornos.

Un estudio en detalle que incluye experimentación y un desarrollo a fondo del algoritmo será publicado próximamente.

## Referencias

1. N. M. Edan, A. Al-Sherbaz, and S. Turner. Design and evaluation of browser-to-browser video conferencing in webrtc. In *2017 Global Information Infrastructure and Networking Symposium (GIIS)*, pages 75–78, Oct 2017.
2. Cheng-Yuan Ho, Fu-Yu Wang, Chien-Chao Tseng, and Ying-Dar Lin. Nat-compatibility testbed: An environment to automatically verify direct connection rate. *Communications Letters, IEEE*, 15(1):4–6, January 2011.
3. C. Jennings and F. Audet. Network address translation (nat) behavioral requirements for unicast udp (rfc 4787). *Network*, 2007.
4. Derek MacDonald and Bruce Lowekamp. Nat behavior discovery using session traversal utilities for nat (stun). Technical report, 2010.

5. C. Medina-López, V. González-Ruiz, and L. G. Casado. On mitigating pollution and free-riding attacks by shamir's secret sharing in fully connected p2p systems. In *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 711–716. IEEE, June 2017.
6. A. Muller, G. Carle, and A. Klenk. Behavior and classification of NAT devices and implications for NAT traversal. *Network, IEEE*, 22(5):14–19, September 2008.
7. A. Muller, N. Evans, C. Grothoff, and S. Kamkar. Autonomous nat traversal. In *International Conference on Peer-to-Peer Computing (P2P)*, pages 1–4. IEEE, August 2010.
8. J. Rosenberg, R. Mahy, P. Matthews, and D. Wing. Session traversal utilities for nat (stun). https://tools.ietf.org/html/rfc5389, October 2008.
9. J. Rosenberg, J. Weinberger, C. Huitema, and R. Mahy. Simple traversal of user datagram protocol through network address translation stun. Technical report, RFC-3489, 2003.
10. D. Senie. Network Address Translator (NAT) - friendly application design guidelines. 2002.
11. P. Srisuresh and M. Holdrege. Ip network address translator (nat) terminology and considerations. 1999.
12. Y. Takeda. Symmetric nat traversal using stun. https://tools.ietf.org/id/draft-takeda-symmetric-nat-traversal-00.txt, June 2003.

# Una arquitectura de microservicios para componentes digitales en el marco del Internet de las Cosas

Manel Mena

Grupo de Investigación de Informática Aplicada (TIC-211), Departamento de Infomática, Universidad de Almería, Ctra. Sacramento S/N, Almería, España
manel.mena@ual.es

**Resumen** La comunicación entre dispositivos del Internet of Things (IoT) es muy heterogénea debido a que estos cuentan con distintos protocolos y servicios para la comunicación con el exterior. Debido a ello, surgen los problemas de la interoperabilidad e integración entre dispositivos o plataformas. Por otro lado, debido a las restricciones que encontramos en la mayoría de los dispositivos IoT (bajo consumo energético/bajo poder de computación), es común encontrar cuellos de botella en la comunicación con este tipo de dispositivos, tanto a la hora de acceder al estado como para poder actuar sobre ellos. Para solucionar estos problemas, proponemos la implementación de una arquitectura de microservicios para la gestión de lo que denominaremos "Digital Quads" (DQ). Los Digital Quads son una representación virtual de dispositivos IoT. Los DQ pretenden dar una solución al problema de la interoperabilidad y el escalado de dispositivos IoT mediante una aproximación holística al mismo. Estos elementos proporcionarán una solución que permita la gestión de eventos y un control de entrada/salida sobre dispositivos IoT utilizando tecnologías web, por lo que nos abstraemos de los protocolos utilizados. Por último, pretendemos hacerlos compatibles con los estándares de la Web of Things (WoT) y prepararlos para que formen parte de un sistema Open Data.

**Palabras Clave:** IoT · Microservicios · WoT · Interoperabilidad · Digital Twin · Protocolos · Open Data

## 1. Introducción

Cuando hablamos de establecer un ecosistema de dispositivos de Internet of Things (IoT), nos encontramos con la problemática de la ingente cantidad de protocolos que los gobiernan. Debido a ello, surge el problema de la interoperabilidad entre dispositivos o plataformas. Para comprender un poco la cantidad de protocolos que existen es común utilizar la siguiente división [1]:

a) **Infraestructura** (e.j., 6LowPAN, IPv4/IPv6, RPL)
b) **Identificación** (e.j., EPC, uCode, IPv6, URIs)

c) **Comunicación / Transporte** (e.j., Wifi, Bluetooth, LPWAN)
d) **Descubrimiento** (e.j., Physical Web, mDNS, DNS-SD)
e) **Protocolos de datos** (e.j., MQTT, CoAP, AMQP, Websocket, Node)
f) **Gestión de dispositivos** (e.j., TR-069, OMA-DM)
g) **Semánticos** (e.j., JSON-LD, Web Thing Model)
h) **Frameworks Multi-capa** (e.j., Alljoyn, IoTivity, Weave, Homekit)

Por otro lado, existen ciertas infraestructuras que intentan dar soporte a la integración de dispositivos IoT en entornos de Smart Home, como pueden ser HomeAssistant, OpenHab, Prodea, etc. Este tipo de infraestructuras están diseñadas para controlar un número controlado de dispositivos, y son lo que se denomina aplicaciones monolíticas, por lo que carecen de una buena capacidad escalabilidad.

Otra problemática que es muy común encontrarnos cuando trabajamos con dispositivos IoT es que, debido a las restricciones en la mayoría de los dispositivos IoT (bajo consumo energético / bajo poder de computación), es común encontrar cuellos de botella en la comunicación, tanto a la hora de acceder al estado como de actuar sobre ellos.

Por último, también surge la necesidad de poder virtualizar este tipo de dispositivos a efectos de poder realizar pruebas sin que esto influya en nuestros procesos de negocio [2]. Para solucionar este último problema surge el concepto de componente virtual o Digital Twin [3] (DT, gemelo digital). Los gemelos digitales son representaciones virtuales de elementos físicos, sistemas o dispositivos a lo largo de su ciclo de vida, que pueden ser usados para diversos propósitos. El concepto de DT, sin embargo, se centra en una aproximación monolítica para la gestión y representación de los dispositivos del mundo real y carecen de una aproximación a múltiples niveles que aborden de forma específica cada una de las facetas de un dispositivo.

Para solucionar todos los problemas planteados anteriormente, proponemos el concepto de Digital Quad (DQ). Al igual que los gemelos digitales, los Digital Quads son representaciones virtuales de elementos físicos, en este nuestro caso dispositivos IoT, pero van más allá. Los Digital Quad proponen una abstracción virtual de dispositivos IoT, esta abstracción estará basada en microservicios y pretende ser agnóstica a los protocolos utilizados por cada dispositivo IoT. En una primera aproximación trabajaremos en la integración de lo que antes hemos denominado protocolos de datos y semánticos.

El concepto de Quad (Quadruplets, cuatrillizos) viene dado por cómo se caracterizan los microservicios que representarán nuestros dispositivos, estos microservicios contarán con diferentes facetas. Como primerá aproximación hemos determinado la necesidad de que los DQ tengan normalmente cuatro facetas, de ahí el nombre de Quads, estas son:

a) **User Interface (UI)**. Esta faceta está compuesta por una representación visual en forma de componente de interfaz de usuario del dispositivo IoT en concreto. Por lo tanto, cuando el usuario requiera tan solo una representación gráfica directa del estado del dispositivo utilizará esta faceta.

b) **Controller**. Esta faceta se compone de las funciones necesarias para operar sobre el dispositivo IoT en cuestión. Algunos de los métodos o rutas de esta faceta requerirán autenticación a nivel de aplicación, a nivel de usuario o a ambos, es decir que el usuario solo pueda acceder a esas rutas si tiene permisos y por una aplicación en concreto.
c) **Data Handler**. Esta faceta se encargará de registrar cualquier tipo de interacción o cambio de estado que se produzca en el dispositivo IoT. Es decir, si estamos hablando de dispositivos actuadores este servicio registrará tanto el cambio de estado del actuador propiamente dicho, como quién ha realizado la interacción sobre el mismo. En el caso de sensores, registrará todo cambio de estado en el sensor, así como quién lo ha consultado y en qué momento para su posterior análisis. A parte de lo expuesto, esta faceta permitirá consultar tanto el estado actual del dispositivo como un histórico de los cambios de estados del mismo, todo ello soportado por una base de datos documental, en nuestro caso MongoDB, que nos ofrece un gran poder de escalabilidad.
d) **Voice Assistant Interface (VAI)**. Esta faceta actuará de enlace con dispositivos de asistencia de voz. Pretendemos estudiar la posibilidad de interconectar nuestros dispositivos IoT con alguno, sino múltiples de los servicios asistentes de voz existentes (Alexa, Google, Siri, etc.).

La Figura 1 representa la composición de nuestros Digital Quads, en ella se aprecia cómo las facetas se intercomunican entre sí y a la vez pueden o no conectarse con el dispositivo IoT que representan. La conexión de las facetas con los dispositivos IoT, es uno de los objetos de estudio en este trabajo de investigación. Pretendemos clasificar el tipo de conexión que se produzca con los dispositivos IoT como fuertemente conectadas o débilmente conectadas. La Figura 1 representa una posible configuración entre dichas facetas y un dispositivo IoT. En la figura apreciamos cómo el Controller puede actuar sobre el dispositivo IoT, mientras que el Data Handler recibe los tanto los datos que se generan en el dispositivo como las interacciones registradas sobre el Controller, en la figura también representamos como las interfaces (UI - VAI) actúan sobre el Controller y reciben los datos a tráves del Data Handler. En el caso concreto del DQ de la figura todas las facetas están fuertemente conectadas, dado que esta representa un actuador en el cual tenemos control total sobre toda su funcionalidad, y a la vez tanto la UI como VAI, son capaces de mostrar datos manejados por el Data Handler e incidir sobre el estado del actuador que es manejado por el Controller. En el caso de que estuviesen debilmente conectadas la intenracción entre facetas es parcial, por ejemplo un controlador puede ofrecer multitud de metodos mientras que la interfaz solo utiliza un subconjunto de ellos.

La posibilidad que tenemos de replicar las facetas que componen nuestros Digital Quads es lo que proporciona la capacidad de escalar el límite máximo de peticiones recibidas sobre nuestros dispositivos IoT. Además, tendremos la posibilidad de escalar tan solo las facetas que se vean más afectadas con respecto al número de peticiones. Por ejemplo, si la faceta que más peticiones está recibiendo es el Data Handler, debido a que uno o varios sistemas de Open Data están realizando tareas de actualización a la vez que muchos usuarios están in-

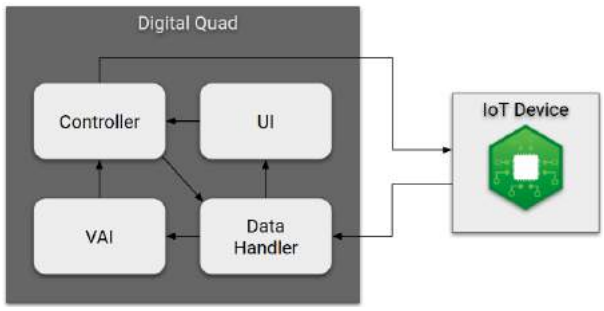



**Figura 1.** Digital Quad

tentando acceder a datos históricos del dispositivo, tendremos la posibilidad de solo replicar esta faceta.

Otro de los pilares sobre los que se sustenta el desarrollo de la tesis es la inclusión y el manejo de información contextual derivadas de los dispositivos IoT. Por ejemplo, la posición espacial (geolocalización) de los mismos puede incidir en la disponibilidad del Digital Quad que lo representa en una región en concreto.

Las facetas de los Digital Quads se comunicarán con el exterior siguiendo los estándares, mecanismos o tecnologías establecidos en el marco de la Web de las Cosas (WoT, Web of Things) [4]. El concepto WoT es una aproximación del IoT que utiliza la tecnología web para acercar a los desarrolladores a este tipo de sistemas embebidos añadiendo una capa de abstracción basada en protocolos web de comunicación e interacción. El proyecto también maneja la posibilidad de lidiar con procesado de eventos complejos [5] (CEP). Estos eventos serán divididos en dos niveles de aplicación:

a) **Individual**. A nivel de un Digital Quad. Por ejemplo, mandar un aviso de un evento de detección de movimiento solo cuando se detecte movimiento en un sensor concreto dos veces en un intervalo de 5 segundos. De esta manera disminuiríamos los falsos positivos.
b) **Múltiple**. A nivel de dos o más Digital Quads. Por ejemplo, disparar un evento de encendido del aire acondicionado cuando se detecte la apertura de la puerta principal de la casa y se encienda la luz de entrada a la casa. Con este tipo de eventos favorecemos la gestión de eventos complejos a nivel de múltiples dispositivos IoT.

Nuestra propuesta tendrá en cuenta la entrada de eventos, su procesamiento y la generación de la respuesta de un Digital Quad de forma individual (aunque formando parte la arquitectura de microservicios propuesta) y cuestiones relativas a su orquestación se dejarán posteriores trabajos de investigación.

Por último, queremos establecer la posibilidad de que nuestros Digital Quads tengan la capacidad de ofrecer a sistemas de Open Data [6] la información que estos requieran, por lo que intentaremos hacerlos compatibles con sistemas del

estilo CKAN, DKAN, OpenDataSoft, etc. Queremos que nuestro propio sistema tenga las características de un sistema Open Data, por ejemplo, en cuestiones relativas a formatos, almacenamiento, estrategias de explotación de la información, frecuencia de actualización de los datos, etc.

La arquitectura de microservicios [7] que proponemos para la gestión de nuestros Digital Quads se muestra en la Figura 2. Esta figura establece una posible configuración propuesta para la arquitectura, donde contaremos con .$^{Ed}$ge"que estará compuesta básicamente de dos tipos de microservicios. Por un lado las Gateway (puertas de enlace) que se encargarán de redirigir las peticiones pertinentes a nuestros Digital Quads, y por otro lado los Discovery Services (servicios de descubrimiento), donde cada vez que se active alguna instancia de servicios contenidos en nuestros Digital Quads queden registrados en los Discovery Services y así nuestros Gateways tendrán la capacidad de realizar balanceo de carga si así se considera oportuno. En segundo lugar, contamos con el Çore"de nuestro sistema, que es donde se enmarcan nuestros Digital Quads y una serie de microservicios auxiliares como pueden ser servicios de autenticación o microservicios de CEP. Además, contaremos con una capa de persistencia, "Persistence", que se compondrá de bases de datos y posibles servicios asosciados a las mismas. Por último, nuestra propuesta contiene lo que denominamos "Things"que se componen de los dispositivos físicos asociados a nuestros Digital Quads, así como servicios externos, componentes virtuales, etc. En la Figura 2 se ejemplifica la gestión de tres Digital Quads. El DQ#1 tiene las cuatro facetas descritas y se encarga de la gestión de un actuador para encender y apagar un interruptor. El DQ#2 solo tiene activas dos facetas debido a que se ha determinado que el servicio de información climatológica que gestiona no debe proporcionar ni UI ni VAI para su acceso, sino que su interacción se realizará a través de las facetas DataHandler y Controller. El DQ#3 tiene la faceta DataHandler duplicada debido a que se ha superado un número de accesos determinado para obtener la información proporcionada por un sensor. De esta manera, pretendemos evitar los posibles cuellos de botella y mejorar el rendimiento en el acceso.

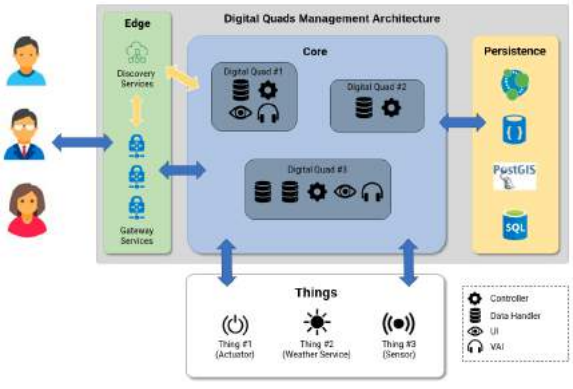



**Figura 2.** Digital Quads Management Architecture

## 2. Avances

Debido a que hemos realizado un cambio en la temática de la tesis doctoral, los avances realizados no han sido muy significativos, puesto que los esfuerzos de investigación se han centrado en el análisis y eleboración de esta propuesta. No obstante, debido a la existencia de nexos comunes entre la línea de tesis anterior y a la que se refiere este documento, en especial el uso de dispositivos IoT restringiéndolos a información contextual (georeferencia) y el uso de arquitecturas de microservicios, podemos enmarcar dos trabajos realizados anteriormente con la nueva propuesta, los cuales son los siguientes:

1. A First Approach towards Storage and Query Processing of Big Spatial Networks in Scalable and Distributed Systems [8].
2. A Progressive Web Application based on microservices involving geospatial data and the Internet of Things [9].

En el primer aporte establecimos una propuesta de cómo gestionar grandes volúmenes de datos espaciales, proporcionando así una manera de tratar nuestros datos contextuales espaciales. En el segundo, vemos cómo existe una estrecha relación con la línea establecida en la nueva tesis doctoral, dado que en ese artículo establecemos una primera aproximación a los dispositivos IoT y lo que es aún más significativo, trabajamos con una propuesta muy parecida a la establecida en la tesis a la hora de definir la arquitectura subyacente en nuestro sistema.

## 3. Conclusiones

Después de realizar la exposición de los problemas y las posibles soluciones dadas en la sección de introducción, nos encontramos con una serie de conceptos en los que la temática de tesis pretende incidir. Buscamos ofrecer una respuesta que permita mejorar interoperabilidad, integración y facilitar la gestión entre sistemas y dispositivos IoT tanto reales como virtuales. Para ello, se llevará acabo la abstracción de la funcionalidad de los dispositivos IoT a un conjunto de microservicios todo ello siguiendo los estándares marcados por la WoT.

El uso de arquitecturas de microservicios trae consigo la posibilidad de proponer una orquestación de los mismos en los que se tenga en cuenta requisitos deseables para nuestro sistema. Estos requisitos a tener en cuenta pueden ser: Escalabilidad de granularidad fina, que nos permite realizar un mejor aprovechamiento de recursos; Un mantenimiento más simple y barato, dado que podemos trabajar mejorando facetas de los Digital Quads de manera individual; La modularidad que tienen nuestros Digital Quads permite una evolución de nuestro sistema de manera mucho más natural, permitiendo mejorar módulo a módulo, e ir extendiendo esa capacidad de gestión de distintos dispositivos poco a poco. Por otro lado, la creación del del concepto de Digital Quad, hace posible controlar dispositivos IoT que requieran de distintas facetas que los representen, por ejemplo, puede haber dispositivos que necesitemos controlar, a traves de servicios web, pero no requieran una representación gráfica de los mismos.

En el proyecto también analizamos la posibilidad de implementar sistemas de CEP tanto a nivel de múltiples Digital Quads, como a nivel interno de cada elemento si se da una secuencia de eventos que deseemos controlar. Por último, planteamos la posibilidad de que nuestro sistema sea compatible con algunos estándares de Open Data como son CKAN, DKAN y OpenDataSoft.

Debido al incremento en el uso de dispositivos y sistemas basados en el concepto de la Web of Things, consideramos que las propuestas que aportamos en la línea de esta tesis pueden ser significativas tanto para la comunidad científica, como la sociedad en general.

## Referencias


1. Postscapes.: IoT Standards & Protocols Guide 2019 Comparisons on Network, Wireless Comms, Security, Industrial. https://www.postscapes.com/internet-of-things-protocols/. Last accessed 4 Feb 2019
2. Shetty, Sony.: How to Use Digital Twins in Your IoT Strategy. Gartner IT Glossary. Gartner IT Glossary. Gartner, Inc. https://www.gartner.com/smarterwithgartner/how-to-use-digital-twins-in-your-iot-strategy/. Last accessed 5 Feb 2019
3. Tuegel, E. J., Ingraffea, A. R., Eason, T. G., Spottswood, S. M.: Reengineering Aircraft Structural Life Prediction Using a Digital Twin. International Journal of Aerospace Engineering, vol. 2011, Article ID 154798, 14 pages. 2011. https://doi.org/10.1155/2011/154798
4. Guinard, D., Trifa, V.: Building the web of things: with examples in node. js and raspberry pi. Manning Publications Co. 2016.
5. Cugola, G., Margara, A.: Processing flows of information: From data stream to complex event processing. ACM Computing Surveys (CSUR), 44(3), 15:1-15:62. 2012. https://doi.org/10.1145/2187671.2187677
6. Ahlgren, B., Hidell, M., Ngai, E. C. H.: Internet of things for smart cities: Interoperability and open data. IEEE Internet Computing, 20(6), 52-56. 2016. https://doi.org/10.1109/MIC.2016.124
7. Krylovskiy, A., Jahn, M., Patti, E.: Designing a smart city internet of things platform with microservice architecture. In 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud), pp. 25-30. IEEE. 2015.
8. Mena, M., Corral, A., Iribarne, L.: A First Approach towards Storage and Query Processing of Big Spatial Networks in Scalable and Distributed Systems. In: XXIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2018), Sevilla (2018). http://hdl.handle.net/11705/JISBD/2018/038
9. Mena, M., Corral, A., Iribarne, L., Criado, J.: A Progressive Web Application based on microservices involving geospatial data and the Internet of Things (under review). Future Generation Computer Systems. 2019.

# Aplicación del IoT en la agricultura intensiva protegida

Manuel Muñoz Rodríguez

[1]Centro Mixto CIESOL, ceiA3, Departamento de Informática, Universidad de Almería.

**Resumen** La agricultura convencional esta sufriendo una serie de cambios debido al aumento en la demanda de productos, precios cada día más ajustados y la escasez de recursos. En consecuencia, esta debe prepararse para enfrentarse a estos desafíos y el uso de la tecnología se puede presentar como uno de los mecanismos para hacer frente a estos nuevos retos. La PA (Precision Agriculture - Agricultura de Precisión) puede ayudar en esta tarea. Se presenta como un conjunto de herramientas TIC que permiten, a través de sensores y otros dispotivos, poder optimizar la gestión de recursos, incrementar la producción y ser mas respetuosos con el medio ambiente. Cabe destacar los beneficios económicos que se derivan de esta, ya que consigue un ahorro en el consumo de recursos como puede ser el agua y un mejor reporte económico al aumentar la producción y disminuir los costes. El objetivo principal de esta tesis doctoral es el desarrollo de sistemas de ayuda a la toma de decisiones en cultivos de invernadero basados en tecnologías relacionadas con el IoT (Internet of Things - Internet de las Cosas), lo que nos obliga a realizar ciertas investigaciones en el ámbito del modelado de datos en IoT, interconectividad entre diferentes servicios e integración de modelos predictivos DSS (Decision Support Systems - Sistemas de ayuda a la toma de decisiones) para ayuda en la toma de decisiones.

**Keywords:** PA · TIC · IoT · DSS.

## 1. Introducción

La FAO (Organización de las Naciones Unidas para la alimentación y agricultura) prevé en 2050 la necesidad de aumentar en más de un 70 % la producción de alimentos de que disponemos hoy en día, lo que se traduce en un importante desafío para la agricultura tal y como la conocemos hasta el momento, debido a los problemas asociados, como son la limitación de tierras cultivadas, el aumento del consumo de agua dulce y el novedoso cambio climático que puede afectar de una forma muy drástica en las formas actuales de producción con la aparición de nuevas enfermedades y plagas hasta ahora desconocidas. Este reto supone un gran cambio para el cual la agricultura debe adaptarse [1], siendo el uso de tecnologías una de las posibles soluciones. La denominada Agricultura Inteligente

(Smart Farming) [2] representa la integración de nuevas tecnologías en el campo de la agricultura, facilitando la llamada "tercera revolución verde"[3]. Esta revolución se basa en la combinación de herramientas IoT [4], procesamiento de datos en Big Data [5], computación en la nube (cloud computing) [6], inteligencia artificial y aprendizaje profundo (IA and deep learning), etc. El uso de estas tecnologías busca lograr un uso eficiente y sostenible de los recursos disponibles para impulsar la producción, incluida la optimización del uso de agua y energía o la prevención contra plagas y enfermedades. La agricultura moderna se basa en una serie de herramientas para ayudar en la mejora de la toma de decisiones, actividades de monitoreo de cultivos y herramientas para mejorar la calidad, producción y cantidad de alimentos. Es un sistema en constante evolución.

Para ayudar al desarrollo de la agricultura inteligente, hemos de recurrir a una serie de sensores, actuadores y dispositivos capaces de proporcionar una información de contexto que tiene su punto de origen en el propio invernadero, generando una gran cantidad de datos [7]. Estos son necesarios para alimentar las diferentes herramientas TIC mencionadas anteriormente (ver Fig 1).


**Figura 1.** Sistema IoT (Hispatec)

Uno de los principales objetivos es ayudar a solucionar el problema de interoperabilidad que existe entre diferentes proveedores de servicios, lo que significa que cada empresa encargada de la venta de sensores y actuadores dispone de su propia plataforma cerrada e inaccesible para la consulta de los datos, lo que plantea un problema dada la necesidad de consultar varias plataforma para obtener la información de un único invernadero y además evitando la interconexión entre servicios, lo cual sería un punto fuerte a la hora de complementar las carencias de un sistema con otro.

Otro de los objetivos es integrar un DSS basado en los datos de contexto generados por los diferentes sensores y actuadores, proporcionando una serie de consignas las cuales ayudan en la producción, crecimiento del cultivo o reducción del consumo de recursos.

## 2. Materiales y Métodos

Esta tesis esta siendo desarrollado dentro del marco de trabajo del proyecto Europeo IoF2020 (Internet of Food and Farm 2020), para su desarrollo, se están obteniendo datos de diferentes sistemas de adquisición como: estaciones meteorológicas, sensores de campo, cuadernos de campo, sistemas de adquisición de datos y control (Supervisory Control And Data Acquisition, SCADA), etc. En el ámbito del proyecto se está colaborando con 7 agricultores con más de 20 sensores, distribuidos por las diferentes zonas geográficas de Almería y diferentes cooperativas, además de la Estación Experimental de la Fundación Cajamar en El Ejido, Provincia de Almería (ver Fig. 2), España (2 43 '00' 'W, 36 48' 00 " N, y 151 m snm), la cual dispone de más de 90 sensores y actuadores conectados a un SCADA.

**Figura 2.** Invernadero utilizado en los ensayos, abajo los cultivos ms representativos de la zona: tomate y pimiento.

Los parámetros climáticos dentro del invernadero fueron monitorizados continuamente cada 30 segundos, este funcionamiento es gestionado por diferentes ETL (Extract, Transform and Load - Extraer, Transformar y Cargar) y procesos Cron encargados de realizar la toma de datos (ver Fig. 3). Fuera del invernadero, una estación meteorológica midió la temperatura del aire y la humedad relativa

con un sensor ventilado, radiación solar, radiación fotosintética activa con un sensor de silicona, detector de lluvia, concentración de $CO_2$, dirección del viento y velocidad del viento.

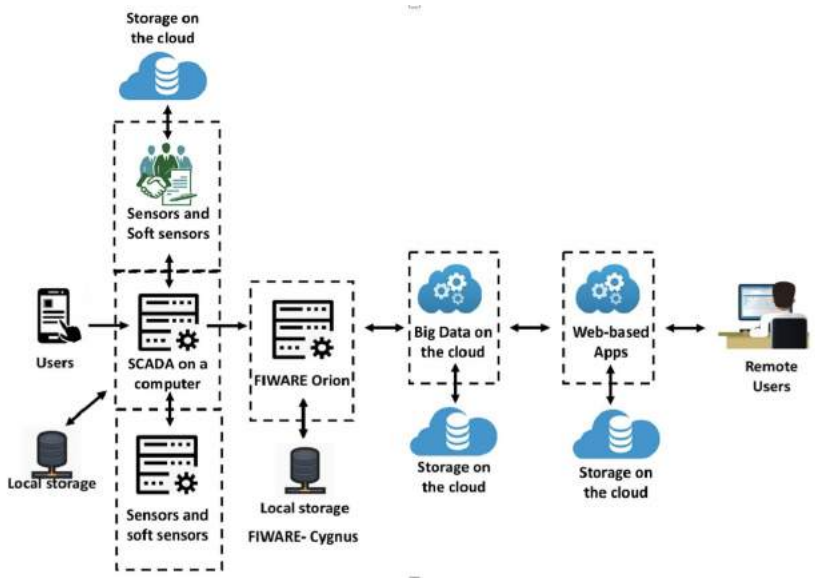



**Figura 3.** Esquema de la plataforma IoT para soporte de decisiones en cultivos en invernaderos.

## 3. Objetivos

El objetivo principal de esta tesis doctoral es el desarrollo de sistemas de ayuda a la toma de decisiones en cultivos de invernadero basados en tecnologías relacionadas con el IoT. Este objetivo conlleva investigaciones en el ámbito del modelado de datos en IoT, interconectividad entre diferentes servicios e integración de modelos predictivos para ayuda en la toma de decisiones. Como se ha indicado, la tesis doctoral se circunscribe en este reto, pudiéndose distinguir los siguientes subobjetivos para cubrir el objetivo principal:

1. Análisis del estado tecnológico en la aplicación de las tecnologías relacionadas con el IoT en la producción de cultivo en invernadero. Estudio de mejoras que se pueden ofrecer respecto al estado actual.
2. Análisis y desarrollo de un modelo de datos para la implementación del paradigma del IoT en agricultura intensiva.
3. Estudio y diseño de una arquitectura IoT para su aplicación en agricultura protegida.

4. Desarrollo de una plataforma tecnológica IoT que facilite la toma de decisiones del agricultor y garantice la interoperabilidad e interconexiones de diferentes proveedores.
5. Desarrollo de algoritmos que permitan fusionar datos heterogéneos, modelos y sistemas de predicción para optimizar distintas funciones de coste relacionadas con la producción, teniendo en cuenta las distintas escalas temporales asociadas al problema. Integración de modelos de clima, riego y crecimiento de cultivo.
6. Establecimiento de un sistema de alertas.
7. Realización de ensayos experimentales para validar la tecnología desarrollada.

## 4. Estado de desarrollo

Este apartado resume los avances en el desarrollo de la tesis. Se han realizado aportes y contribuciones en algunos de los objetivos mencionados anteriormente, los cuales serán descritos a continuación.

- *Estudio y diseño de una arquitectura IoT para su aplicación en agricultura protegida.* Se ha desarrollado una primera fase de un modelo de datos estándar para IoT. Este modelo se basó en bases de datos relacionales y arquitecturas cloud como Fiware para la integración de los datos. Esta información fue representada en el articulo [11] en el congreso III Symposium Nacional de Ingeniería Hortícola en Lugo. Este articulo recibió el premio de la Sociedad Española de Agroingeniería a la mejor aportación.
- *Análisis y desarrollo de un modelo de datos para la implementación del paradigma del IoT en agricultura intensiva.* Para la estandarización en el modelo de datos de comunicación se utiliza la plataforma desarrollada por la Unión Europea Fiware, la cual trata de imponerse con estándares propios unificando un conjunto de API'S. Esta información fue representada en el articulo [12] en el congreso AGENG 2018 - EURAGENG CONFERENCE en Wageningen.
- *Desarrollo de una plataforma tecnológica IoT que facilite la toma de decisiones del agricultor y garantice la interoperabilidad e interconexiones de diferentes proveedores.* Se ha desarrollado una primera fase de la aplicación, la cual reúne datos de diferentes proveedores de servicios y los representa a través de un portal web, donde el agricultor puede consultar en tiempo real el estado del invernadero y realizar consultas históricas para comprobar las tendencias del cultivo, este articulo [13] se envió al congreso I Congreso de Jóvenes Investigadores en Ciencias Agroalimentarias en Almería.

## 5. Conclusiones

En la presente tesis doctoral se pretende desarrollar sistemas de ayuda a la toma de decisiones en cultivos de invernadero basados en tecnologías relacionadas con el IoT, interopelabilidad entre diferentes servicios y desarrollo de una

aplicación la cual englobe todo lo mencionado. Se han realizado contribuciones a este sector con un total de 3 artículos en congresos [11,12,13] destinados a mejorar la interoperabilidad entre servicios, la cual ayude a integrar diferentes sistemas en una misma herramienta facilitando al consumidor la administración de su finca.

Para trabajos futuros se integrarán modelos de datos predictivos DSS basados en algoritmos matemáticos, los cuales estimarán el clima en el interior del invernadero según las condiciones externas, al igual que modelos basados en el crecimiento del cultivo.

## Referencias

1. R. R. Shamshiri et al., Advances in greenhouse automation and controlled environment agriculture: A transition to plant factories and urban agriculture, Int. J. Agric. Biol. Eng., vol. 11, no. 1, pp. 122, 2018.
2. J. Liu and J. P. Tao, Research and application of agricultural greenhouse intelligence platform based on IoT (Internet of Things) and cloud computing, Int. J. Simul. - Syst., Science and Tech., vol. 17, no. 5, p. 8.1-8.5, 2016.
3. G. Gomes and A. Pérez García, El proceso de modernización de la agricultura latinoamericana, Rev. CEPAL, 8, 57-77, 1979.
4. J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, Internet of Things (IoT): A vision, architectural elements, and future directions, Futur. Gener. Comput. Syst., vol. 29, no. 7, pp. 16451660, Sep. 2013.
5. A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Bold, A review on the practice of big data analysis in agriculture. Comput. Electron. Agric., 143, 23-37, 2017.
6. K. R. Bidua and C. N. Patel, International journal of innovative and emerging research in engineering internet of things and cloud computing for agriculture in India, Int. J. Innov. Emerg. Res. Eng., vol. 2, no. 12, 2015.
7. V. I. Adamchuk et al., On-the-go soil sensors for precision agriculture, Comput. Electron. Agric., vol. 44, pp. 7191, 2004.
8. M. Fazio, A. Celesti, F. G. Marquez, A. Glikson, and M. Villari, Exploiting the FIWARE cloud platform to develop a remote patient monitoring system, in 2015 IEEE Symp. on Computers and Communication (ISCC), 2015, pp. 264270.
9. P. Rajalakshmi and S. Devi Mahalakshmi, IoT based crop-field monitoring and irrigation automation, in 2016 10th Int. Conf. on Intelligent Systems and Control (ISCO), 2016, pp. 16.
10. B. Moltchanov and O. R. Rocha, A context broker to enable future IoT applications and services, in 2014 6th Int. Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2014, pp. 263268.
11. M. Muñoz, J. A. Sanchez, F. Rodriguez, M. Berenguel, C. Giagnocavo, IoT applied to traceability and decision making for greenhouse tomato crops, in III Symposium Nacional de Ingeniería Hortícola, Lugo, España, 2018.
12. M. Muñoz, J. A. Sanchez, F. Rodriguez, M. Berenguel, C. Giagnocavo, Farms, Fogs and Clouds: Data open-architecture for optimal crop growth control for IoF2020 project, in New engineering concepts for a valued agriculture AgEng, Wageningen, The Netherlands, 2018.
13. M. Muñoz, J. A. Sanchez, F. Rodriguez, M. Berenguel, C. Giagnocavo, La importancia del IoT en la agricultura: herramienta para la ayuda a la toma de decisiones,

in I Congreso de Jóvenes Investigadores en Ciencias Agroalimentarias, Almería, España, 2018.

# Análisis de imágenes multi-espectrales aplicadas al campo de la agricultura

Luis Ortega López[1]

Universidad de Almería, lol.teleco@gmail.com

**Resumen** La cámara multi-espectral Sequoia ha sido diseñada especialmente para su uso en el campo de la agricultura de precisión. Sin embargo, las imágenes que se obtienen de cada uno de los cinco sensores de los que dispone presentan una serie de distorsiones geométricas que es preciso corregir antes de trabajar con dichas imágenes. En concreto, en este trabajo se han corregido las imágenes obtenidas por dicha cámara de forma automática. Para ello, se han generado conjuntos de calibración y de testeo que han permitido cuantificar la distorsión de cada una de las cinco lentes de una forma precisa. Una vez determinado y corregido el efecto de distorsión de las lentes, se está trabajando en el preprocesado de imágenes reales de campo, donde se está estudiando la segmentación de las plantas en los diferentes espectros.

## 1. Introducción

El sensor multiespectral Sequoia captura imágenes tanto del espectro visible como del no visible, proporcionando datos para analizar la salud y vigor de las cosechas. Sequoia capta diferentes longitudes de onda, Verde, Roja, Red-Edge (Borde rojo) y NIR (Infrarrojo cercano), para destacar la salud de las plantas. De este modo, es ampliamente utilizada para la agricultura de precisión, que consiste en en el manejo diferenciado de los cultivos a partir del conocimiento de la variabilidad existente en una explotación agrícola [1]. Para que las tareas de análisis de salud y vigor de las plantaciones se realicen de forma exitosa, es necesario que las fotografías tomadas por los cinco sensores sean lo más fiable posible. Por tanto, merece la pena calibrar los distintos sensores de Sequoia ya que las lentes de podrían presentar diferentes tipos de defectos en algunas distancias focales, distancias de enfoque, etc. que dificultarían el proceso de monitorización [2,3].

Sequoia, al tomar una foto, proporciona cinco imágenes independientes (una por cada sensor). Para poder trabajar con dichas imágenes, es necesario superponerlas en un ortomosaico, permitiendo así el posterior reconocimiento digital y el análisis de tales datos. Actualmente, el software PiX4D que viene instalado en la cámara Sequoia, lleva a cabo automáticamente la tarea de generar mapas indexados y de ortomosaicos. Sin embargo, la superposición de las imágenes de cada sensor la lleva a cabo sin tener en cuenta el desplazamiento que existe entre los cinco sensores y, además, no corrige algunas de las aberraciones de la imagen.

En este trabajo, nos hemos centrado en corregir las imágenes obtenidas por la cámara Sequoia de forma automática.

Para ello, hemos considerado conjuntos de calibración y de testeo que han permitido cuantificar la distorsión de cada una de las cinco lentes de forma precisa. Tras corregir los efectos de distorsión de las lentes, se está trabajando en el preprocesado de imágenes reales de campo, donde se está estudiando la segmentación de las regiones de interés (plantas).

## 2. Avances

Los avances se han llevado a cabo en la automatización del proceso de calibrado de cada uno de los 5 sensores de la cámara Sequoia. A continuación, se describe la metodogía que se ha seguido para calibrar cada sensor.

Para cada uno de los cinco sensores se han utilizado unas 120 imágenes aproximadamente. Estas imágenes son de un damero de 28mm de lado por celda tomadas desde varios ángulos. Dado el set de imágenes completo, se han reservado un 20 % de las imágenes para el testeo posterior. Nos quedan, aproximadamente:

- Aproximadamente 100 imágenes para calibración.
- Aproximadamente 20 imágenes para testeo del error.

### 2.1. Calibración mediante el Toolbox de MATLAB

Se han realizado varias iteraciones para calibrar las cámaras mediante el toolbox de Matlab de calibración, eliminando aquellas imágenes que por sus características eran negativas para la calibración. Estas imágenes correspondían a varios tipos:

- Imágenes donde el damero estaba cortado o era irreconocible.
- Imágenes borrosas.
- Imágenes poco representativas que arrojaban mucho error al ser rectificadas con los parámetros calculados. Es decir, los extremos estadísticos que se ajustaban peor a la calibración. Para ello, se han eliminado aquellas imágenes que arrojaban un error medio de más de 0.6 píxeles.

De este modo, se han obtenido 5 conjuntos de parámetros de calibración, uno por cámara.

### 2.2. Testeo del error con los sets de testeo

Para el testeo del error se ha elaborado una rutina en Matlab. Utilizando tal rutina, se describe el proceso que se ha llevado a cabo a continuación:

1. Se cargan las imágenes de un set de testeo, en este caso el GRE. Las imágenes que no son válidas (bien porque no podemos detectar el damero antes de la corrección o después) son eliminadas. El hecho de que no podamos detectar

el damero después de la corrección está relacionado con que es posible que la imagen resultante tras la corrección de la distorsión quede ampliada, y partes del damero se pierdan, haciéndolo irreconocible. Por tanto, dichas imágenes se han detectado y eliminado. La Figura 1 muestra las imágenes válidas para testear el sensor GRE.

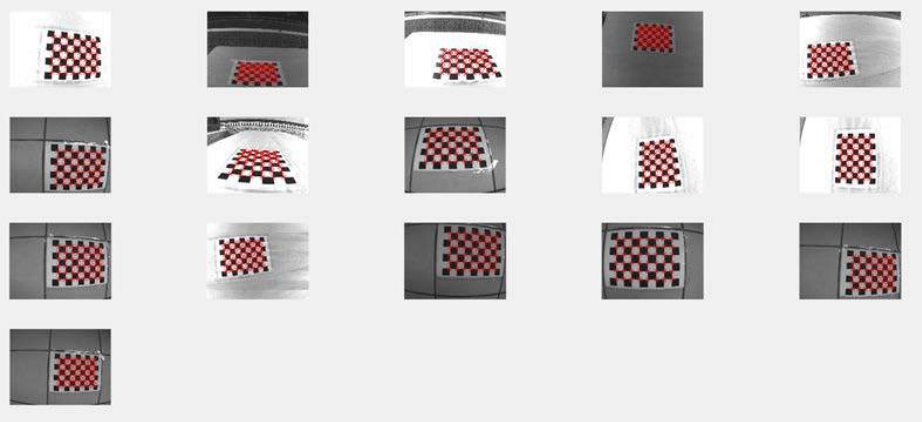


**Figura 1.** Ejemplo de detección de intersecciones para el sensor GRE. En dicho sensor, partimos de 20 imágenes, pero hemos procesado sólo 16 (tras quitar las imágenes no válidas). Sobre estas imágenes no tenemos problema en detectar las 48 intersecciones de las celdas.

2. Se corrige la distorsión en las imágenes utilizando los parámetros que hemos calculado para esa lente en concreto. Obtenemos las 4 intersecciones exteriores sobre las que vamos a rectificar la imagen de forma automática, tal y como se muestra en la Figura 2.
3. Se rectifica la imagen y se trae al plano principal. Sobre esta imagen rectificada podemos calcular donde deberían estar las intersecciones idealmente. Recogemos estas coordenadas ideales y las guardamos en la matriz $A$. Este proceso se muestra en la Figura 3.
4. Se vuelven a reconocer las intersecciones en la imagen corregida. Esto nos dice dónde están realmente las intersecciones tras el proceso de corrección. Esta información se almacena en la matriz $B$.
5. Finalmente, calculando el valor absoluto de $A - B$ se ha podido obtener el error por intersección. También se ha calculado la media del error por intersección, lo cual nos proporciona el error por imagen. Para representar el error por imagen de las 16 imágenes se ha utilizado una visualización estadística por percentiles.
6. Finalmente, se ha calculado el error cuadrático medio.

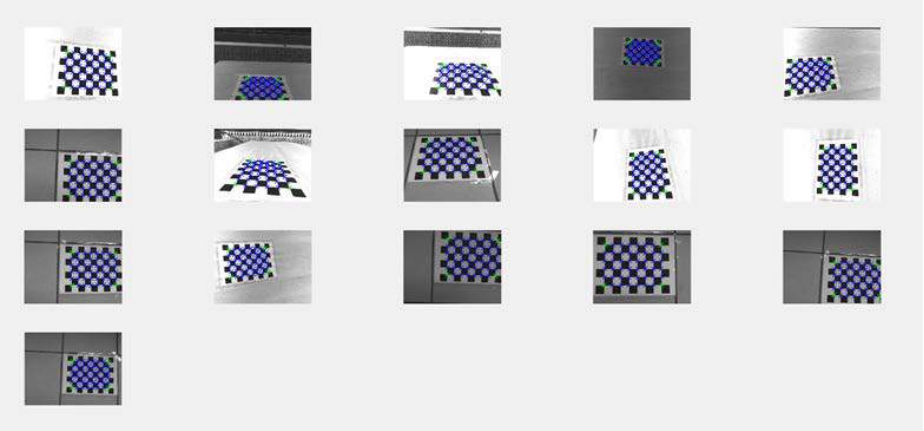


**Figura 2.** Obtención de las 4 intersecciones exteriores para el sensor GRE.

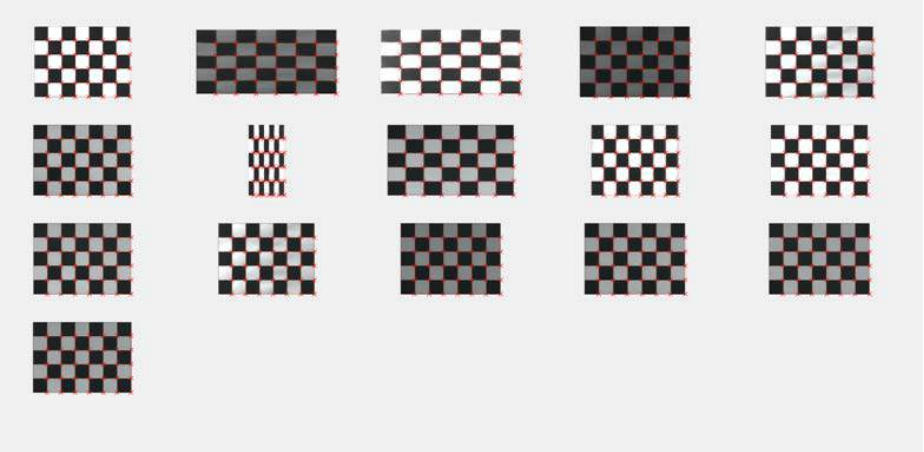


**Figura 3.** Rectificación de las imágenes del sensor GRE y su posterior paso al plano principal.

## 3. Resultados y conclusiones

Los resultados obtenidos en el proceso anteriormente comentado han sido los que se muestra en la Tabla 1. En dicha tabla, la columna $S$ hace referencia a los distintos sensores de Sequoia, la columna $NIT$ se refiere al número de imágenes total de partida, la columna $NICI$ identifica el número de imágenes que se consideraron inicialmente para el set de calibración, $NIC$ es el número de imágenes que se han utilizado realmente para para el set de calibración, $NITI$ es el número de imégenes que se consideraron inicialmente para el set de testeo, $NIT$ identifica el número de imágenes que se han utilizado realmente para el

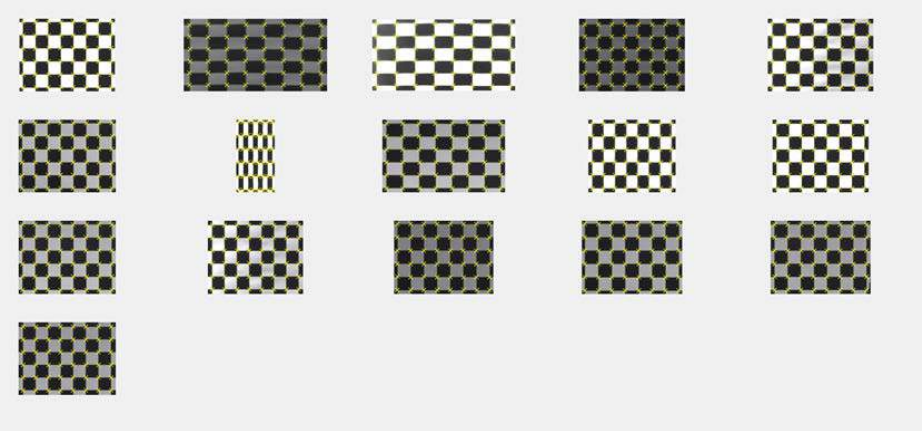


**Figura 4.** Reconocimiento de las intersecciones en las imágenes corregidas (sensor GRE).

testeo tras la calibración de los sensores y $E$ hace referencia al error medio del proceso de calibrado.

**Tabla 1.** Tabla resumen del proceso de calibración.

| $S$ | $NIT$ | $NICI$ | $NIC$ | $NITI$ | $NIT$ | $E$ |
|---|---|---|---|---|---|---|
| GRE | 118 | 98 | 69 | 20 | 16 | 0.2693 |
| NIR | 122 | 100 | 61 | 22 | 14 | 0.2332 |
| RED | 118 | 98 | 78 | 20 | 17 | 0.2984 |
| REG | 115 | 95 | 56 | 20 | 17 | 0.2785 |
| RGB | 123 | 103 | 25 | 20 | 2 | 1.0547 |

Los resultados obtenidos de la Tabla 1 muestran que:

- Sería conveniente obtener más imágenes de testeo y calibración de la lente RGB. Aunque ésta no es de vital importancia para el proceso de vigorosidad de las plantas que queremos llevar a cabo.
- Se ha obtenido una buena magnitud de corrección de la distorsión, pues en todos los espectros de interés se ha obtenido un error medio inferior a 0.3 píxeles.
- El espectro RED es el que más calidad arroja en la cámara. Por el contrario, el espectro REG es el que arroja peor calidad de imagen basándonos en la capacidad de reconocimiento de patrones de los algoritmos utilizados.

Actualmente, estamos trabajando en un proceso de segmentado sobre las imágenes corregidas. Dichas imágenes contienen una planta (región de interés) que está enmarcada por un cuadrado. Por tanto, estamos estudiando como seg-

mentar dicho cuadrado y traerlo al plano principal para, posteriormente, segmentar la planta.

## Referencias

1. S. Fountas et al., "Management strategies and practices for precision agriculture operations," in *Precision Agriculture 2009 - Papers Presented at the 7th European Conference on Precision Agriculture, ECPA 2009*, 2009, pp. 893–898.
2. J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, Washington, DC, USA, 1997, CVPR '97, pp. 1106–, IEEE Computer Society.
3. Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

# Transmisión de secuencias de imágenes JPEG2000 usando actualización condicional y compensación de movimiento controlada por el cliente

José Juan Sánchez Hernández[1]

Universidad de Almería, Ctra. Sacramento, s/n, Almería, 04120. España

**Abstract.** Este trabajo propone una estrategia para la transmisión de secuencias de imágenes JPEG2000 de alta resolución utilizando técnicas de actualización condicional y compensación de movimiento, en arquitecturas cliente/servidor que hagan uso del estándar de compresión JPEG2000 y el protocolo JPIP. Una de las principales contribuciones que se realizan en este trabajo, es que la actualización condicional y la compensación de movimiento se realizan en el lado del cliente, de modo que la técnica propuesta es compatible con cualquier servidor JPIP estándar. Este trabajo aprovecha la escalabilidad espacial que ofrece JPEG2000 para reutilizar los precintos de imágenes que ya han sido reconstruidas y visualizadas previamente por el cliente, y permite determinar qué precintos de la siguiente imagen de la secuencia han cambiado. De modo que el cliente en lugar de solicitar todos los precintos de la siguiente imagen, sólo solicitará aquellos que hayan cambiado y realizará una reconstrucción de la siguiente imagen haciendo uso de los precintos existentes de imágenes previas junto a los nuevos precintos que ha solicitado al servidor. Los resultados de nuestros experimentos demuestran que la calidad de las imágenes reconstruidas mejoran significativamente cuando además de la actualización condicional también se realiza una compensación de movimiento.

**Keywords:** JPEG2000 · JPIP · conditional-replenishment · motion-compensation.

## 1 Introducción

Entre las principales características del estándar JPEG2000 [1] podemos destacar la eficiencia al realizar compresión de datos con pérdida y sin pérdida, el acceso aleatorio a los datos comprimidos, la decodificación incremental de los datos y la alta escalabilidad. Estas características hacen que JPEG2000 sea una solución idónea para la exploración remota de imágenes de alta resolución. En la Parte 9 [2] del estándar se define el protocolo JPIP, que es el protocolo que permite a los clientes explorar datos de imágenes remotas de forma interactiva especificando una ventana de interés (WOI). Este intercambio de datos entre cliente y servidor utiliza el ancho de banda disponible de manera eficiente y no requiere la

recodificación de las imágenes, ni ningún procesamiento adicional. En este proceso de intercambio el servidor extrae sólo los datos requeridos de las imágenes comprimidas y los transmite a los clientes.

En particular, el protocolo JPIP ha demostrado ser muy eficiente para la visualización de imágenes solares. Actualmente está siendo utilizado por el Proyecto JHelioviewer [3] para permitir a los investigadores y al público en general explorar datos de imágenes solares de diferentes observatorios espaciales, hacer zoom de manera interactiva en áreas de interés y reproducir secuencias de imágenes de alta resolución con diferentes niveles de cadencia temporal.

Dependiendo de la resolución de las imágenes, su contenido y la cadencia temporal que el usuario ha seleccionado, la cantidad de datos que deben enviarse desde el servidor hasta el cliente puede ser muy elevada. Este problema es más evidente cuando aumenta el número de usuarios que están realizando peticiones al servidor de forma concurrente. Para resolver este problema de escalabilidad, nosotros proponemos explotar la redundancia temporal que existe en las secuencias de imágenes que se están transmitiendo, realizando compensación de movimiento y refresco condicional.

Aunque el uso de compensación de movimiento y refresco condicional ya han sido utilizados en otros trabajos, nuestra propuesta es la primera en ser compatible al 100% con cualquier servidor JPIP. Esto significa que cualquier servidor JPIP estándar, como el que se está utilizando actualmente en el Proyecto JHelioviewer, puede hacer uso de esta solución. De hecho, nuestra propuesta sólo requiere que los clientes tengan algún grado de acceso aleatorio espacial a las imágenes en el servidor.

## 2 MCCR (Motion Compensated Conditional Replenishment)

Esta sección describe MCCR, una propuesta totalmente compatible con el estándar JPIP [2] que permite la transmisión de secuencias de imágenes JPEG2000 utilizando las técnicas de compensación de movimiento y refresco condicional de precintos JPEG2000. La lógica de MCCR está implementada exclusivamente en el lado del cliente, mientras que el servidor no necesita realizar ningún cambio. Precisamente esta característica es la que hace que esta propuesta sea realmente interesante.

Mediante el protocolo JPIP, los clientes pueden solicitar ventanas de interés (WOIs) a los servidores, que responden con uno o más *data-bins*, que son fragmentos de *codestream* de un archivo JPEG2000. Es importante tener en cuenta que: (1) los clientes solicitan WOIs al servidor, no solicitan paquetes ni *data-bins* y (2) debido a que dos o más WOIs pueden solaparse, los clientes almacenan los paquetes que reciben en una caché local, de modo que los servidores JPIP pueden conocer el estado de las cachés de los clientes con el objetivo de evitar tener que enviar los mismos datos que ya ha enviado previamente a los clientes en la misma sesión JPIP.

### 2.1 Descripción

Para facilitar la descripción de MCCR, vamos a suponer que el cliente va a solicitar la secuencia de imágenes en su máximo nivel de resolución temporal, es decir, no van a existir imágenes intermedias entre $I_i$ and $I_{i+1}$. Sin embargo, MCCR permite utilizar cualquier resolución temporal en el cliente para renderizar la secuencia de imágenes.

A continuación se describe la secuencia de pasos del algoritmo MCCR.

1. En primer lugar se descargan las dos primeras imágenes de la secuencia en su máximo nivel de resolución y con el máximo número de capas de calidad. A estas imágenes las denominamos $I_i$ y $I_{i+1}$, donde $i$ representa en qué iteración del algoritmo nos encontramos.
2. Las imágenes renderizadas que se muestran en el cliente se denominan $\tilde{I}$. Para las dos primeras imágenes de la secuencia tenemos que las imágenes que hemos descargado son las mismas imágenes que se visualizan en el cliente, es decir, $\tilde{I}_i = I_i$ y $\tilde{I}_{i+1} = I_{i+1}$.
3. Realizamos la estimación de movimiento entre las imágenes $\tilde{I}_i$ y $\tilde{I}_{i+1}$ utilizando una precisión subpixel igual a $A$ y un área de búsqueda de $S \times S$ pixels. Para calcular los vectores de movimiento, $\overrightarrow{V}$, utilizamos un estimador de movimiento basado en bloques de tamaño $B \times B$ pixels.
4. Una vez calculados los vectores de movimiento, generamos una imagen predicción $\hat{I}_{i+2}$, proyectando los vectores de movimiento $\overrightarrow{V}$ sobre la imagen $\tilde{I}_{i+1}$. Es decir,

$$\hat{I}_{i+2} = \overrightarrow{V}(\tilde{I}_{i+1}). \tag{1}$$

   Hay que tener en cuenta que que estamos suponiendo que existe un movimiento constante de los bloques entre las imágenes de la secuencia.
5. A partir de la imagen $\hat{I}_{i+2}$ obtenemos una representación en un bajo nivel de resolución, $\texttt{Thumbnail}(\hat{I}_{i+2})$.
6. Solicitamos al servidor una representación en un bajo nivel de resolución de la imagen $I_{i+2}$, $\texttt{Thumbnail}(I_{i+2})$.
7. Calculamos las diferencias a nivel de pixel que existen entre el *thumbnail* de la siguiente imagen de la secuencia, $\texttt{Thumbnail}(I_{i+2})$ y el *thumbnail* de la imagen predicción, $\texttt{Thumbnail}(\hat{I}_{i+2})$.

$$E = \texttt{Thumbnail}(I_{i+2}) - \texttt{Thumbnail}(\hat{I}_{i+2}) \tag{2}$$

8. Calculamos la media del error $E$ que existe en cada una de las WOIs y ordenamos la lista de WOIs en modo descendente. De modo que el resultado contiene una lista $L$ de WOIs que están ordenadas por distorsión. Las WOIs que aparezcan en primer lugar serán las primeras WOIs que deberemos solicitar al servidor.
9. A partir de la lista $L$ y dependiendo del ancho de banda disponible, solicitamos un número determinado de WOIs (con todos los niveles de calidad) de la imagen $I_{i+2}$. Las WOIs seleccionadas serán las que aparezcan en las primeras posiciones de la lista $L$.

10. Creamos la imagen $\tilde{I}_{i+2}$, que será el resultado de mezclar los precintos (WOIs) que hemos recibido de la imagen $I_{i+2}$ con los precintos que teníamos previamente en la imagen $\hat{I}_{i+2}$. La imagen resultado $\tilde{I}_{i+2}$ será la que visualizará en el cliente.
11. $i \leftarrow i + 1$. Ir al paso 4.

### 2.2 Requisitos de compresión de las imágenes

Para poder usar MCCR, es necesario: (1) tener acceso aleatorio sobre los datos comprimidos de la imagen que se corresponden con las WOIs que utilizamos para calcular las predicciones, y (2) las imágenes deben estar comprimidas teniendo en cuenta que deben tener el mismo número de precintos en cada nivel de resolución. También habrá que tener en cuenta que el número $D$ de niveles DWT (Discrete Wavelet Transform) que podemos aplicar sobre las imágenes debe seleccionarse teniendo en cuenta la resolución que queremos que tengan los *thumnails* con los que vamos a trabajar.

## 3 Resultados experimentales

### 3.1 Speedway

La secuencia Speedway es una secuencia de imágenes en escala de grises, con poco movimiento, capturada por una cámara fija de vigilancia a 30 frames/segundo. La secuencia muestra un fondo estático y varios coches circulando por una autopista. La resolución de las imágenes de esta secuencia es de $384 \times 320$. La Figura 1 muestra las 4 primeras imágenes de la secuencia, la primera muestra la imagen original de la secuencia y las tres imágenes siguientes muestran las diferencias que existen en cada imagen respecto a la primera imagen.

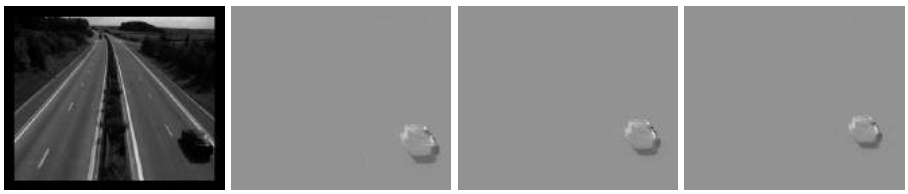


**Fig. 1.** Primeras imágenes de la secuencia Speedway. Se muestra la primera imagen y las diferencias que existen en las imágenes 2, 3 y 4, respecto a la primera imagen de la secuencia.

Las primeras 50 imágenes de la secuencia Speedway han sido comprimidas con JPEG2000, generando 3 niveles de resolución espacial y 8 capas de calidad. El tamaño de code-block es de $16 \times 16$ coeficientes. Los precintos en el máximo

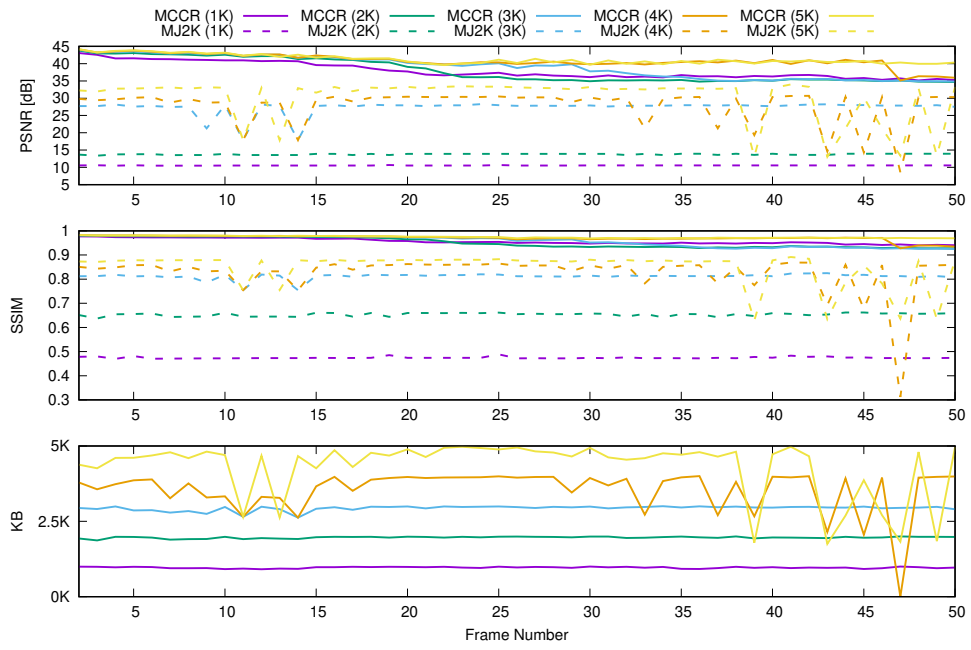



**Fig. 2.** Calidad de las reconstrucciones de las primeras 50 imágenes de la secuencia Speedway utilizando MCCR y Motion JPEG2000 (MJ2K). Se han simulado varios escenarios con diferente ancho de banda donde la cantidad de bytes transmitidos por imagen en cada uno de ellos varía desde los 1000 hasta los 5000 bytes. Se han utilizado dos métricas para evaluar la distorsión de las reconstrucciones: PSNR y SSIM.

nivel de resolución tienen unas dimensiones de $64 \times 64$, y para la estimación de movimiento se han utilizado los siguientes valores $A = 1$ y $S = 4$.

La Figura 2 muestra la calidad de las reconstrucciones que se han obtenido simulando diferentes escenarios con distinto ancho de banda. En cada simulación la cantidad de bytes que se ha transmitido para cada imagen va desde los 1000 bytes hasta los 5000 bytes. Para evaluar la distorsión de las reconstrucciones hemos utilizado dos métricas: PSNR (Peak Signal-to-Noise Ratio) y SSIM (Structural SIMilarity).

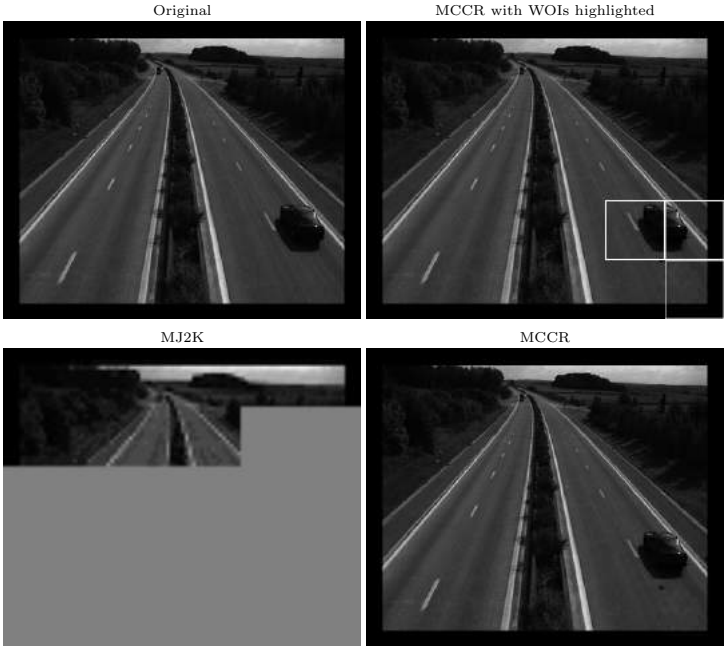



**Fig. 3.** Reconstrucción de la tercera imagen de la secuencia Speedway utilizando MCCR y Motion JPEG2000 cuando se transmiten 1000 bytes por imagen. En la esquina superior izquierda se muestra la imagen original, en la esquina superior derecha está la reconstrucción obtenida con MCCR, se han resaltado las WOIs que se han actualizado. En la esquina inferior derecha tenemos la misma imagen pero sin resaltar las WOIs y en la esquina inferior izquierda está la reconstrucción obtenida con Motion JPEG2000.

La Figura 3 muestra las reconstrucciones que se han obtenido para MCCR y Motion JPEG2000 cuando se transmiten 1000 bytes por imagen de la secuencia Speedway. Como se puede ver, la calidad visual que proporciona MCCR es superior a la que ofrece Motion JPEG2000. De hecho, la calidad de las reconstrucciones obtenidas con MCCR es superior a las obtenidas con Motion JPEG2000

para todos los bit-rates evaluados y según las dos métricas analizadas, PSNR y SSIM.

### 3.2 Stockholm

Stockholm es una secuencia de imágenes a color de $1280 \times 720$, con un nivel de movimiento medio, donde una cámara realiza una panorámica del paisaje de la ciudad de Estocolmo a 50 frames/segundo. En la secuencia aparecen casas, agua y coches en movimiento. La Figura 4 muestra las cuatro primeras imágenes de las secuencia, la primera muestra la imagen original de la secuencia y las tres imágenes siguientes muestran las diferencias que existen en cada imagen respecto a la primera imagen.

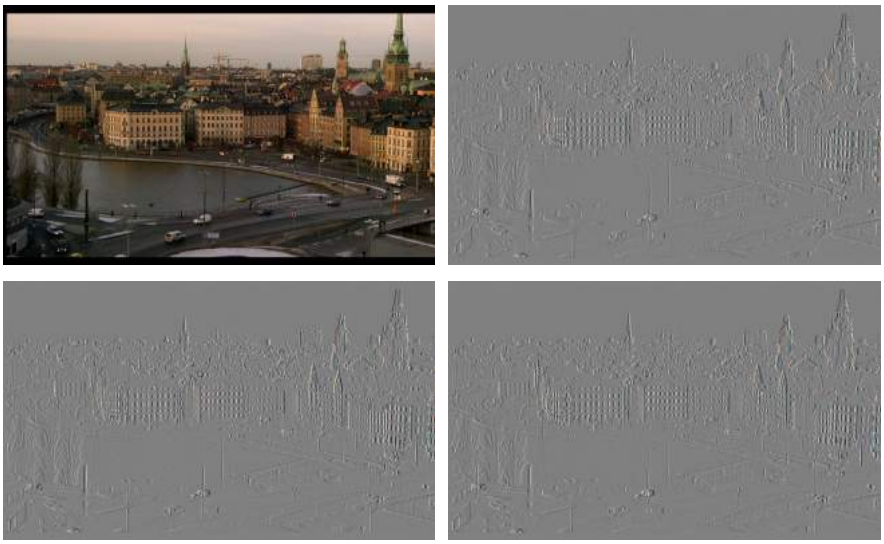


**Fig. 4.** Frames 3, 4, 5 y 6 de la secuencia Stockholm. Se muestra la imagen 3 y las diferencias que existen en las imágenes 4, 5 y 6, respecto a la imagen 3 de la secuencia. Las diferencias muestran que en esta secuencia existe un movimiento mayor que el de la secuencia Speedway.

Las primeras 100 imágenes de la secuencia Stockholm han sido comprimidas con JPEG2000, generando 3 niveles de resolución espacial y 8 capas de calidad. El tamaño de code-block es de $32 \times 32$ coeficientes. Los precintos en el máximo nivel de resolución tienen unas dimensiones de $128 \times 128$, y para la estimación de movimiento se han utilizado los siguientes valores $A = 2$ y $S = 4$.

La Figura 5 muestra la calidad de las reconstrucciones que se han obtenido simulando diferentes escenarios con distinto ancho de banda. En cada simulación la cantidad de bytes que se ha transmitido para cada imagen va desde

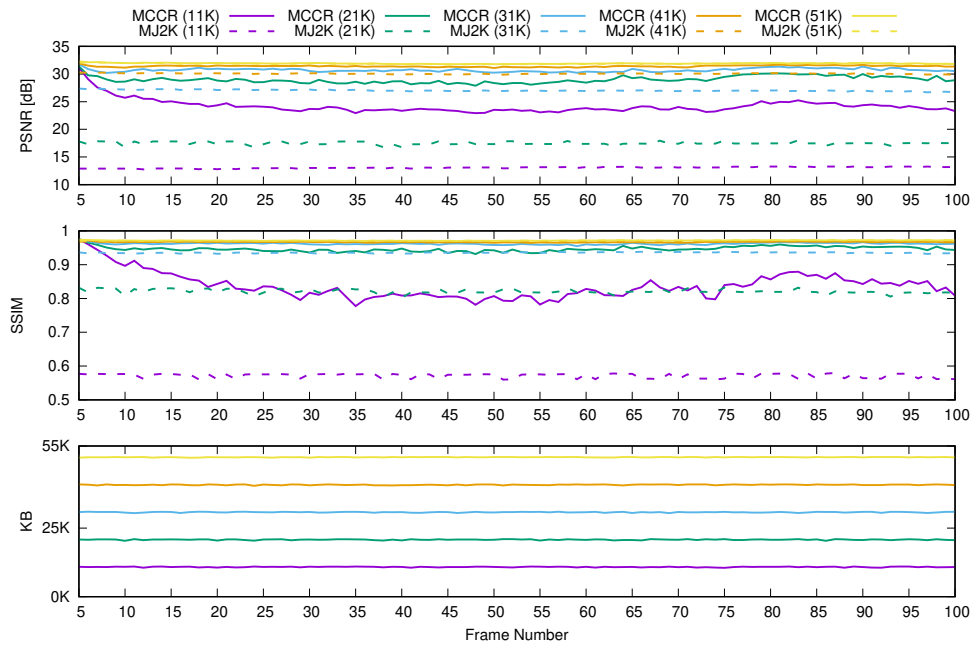



**Fig. 5.** Calidad de las reconstrucciones de las primeras 100 imágenes de la secuencia Stockholm utilizando MCCR y Motion JPEG2000 (MJ2K). Se han simulado varios escenarios con diferente ancho de banda donde la cantidad de bytes transmitidos por imagen en cada uno de ellos varía desde los 11000 hasta los 51000 bytes. Se han utilizado dos métricas para evaluar la distorsión de las reconstrucciones: PSNR y SSIM.

los 11000 bytes hasta los 51000 bytes. En este caso tenemos una secuencia con mayor movimiento que en la secuencia Speedway, y los resultados muestran el comportamiento esperado:

1. Cuanto más complejo es el movimiento de la secuencia, existe menor correlación temporal entre las imágenes de la secuencia y por lo tanto MCCR ofrece menor beneficio en las reconstrucciones.
2. Si el bit-rate de las transmisiones es muy pequeño, entonces el número de WOIs que son actualizadas en la predicción no es suficiente para obtener una buena reconstrucción de las imágenes. Este efecto se puede observar cuando utilizamos un bit-rate de 11000 bytes por imagen y puede ser resuelto monitorizando la calidad en el cliente y solicitando una imagen "intra" cuando la calidad de las reconstrucciones se encuentre por debajo de un umbral que esté controlado por el usuario.

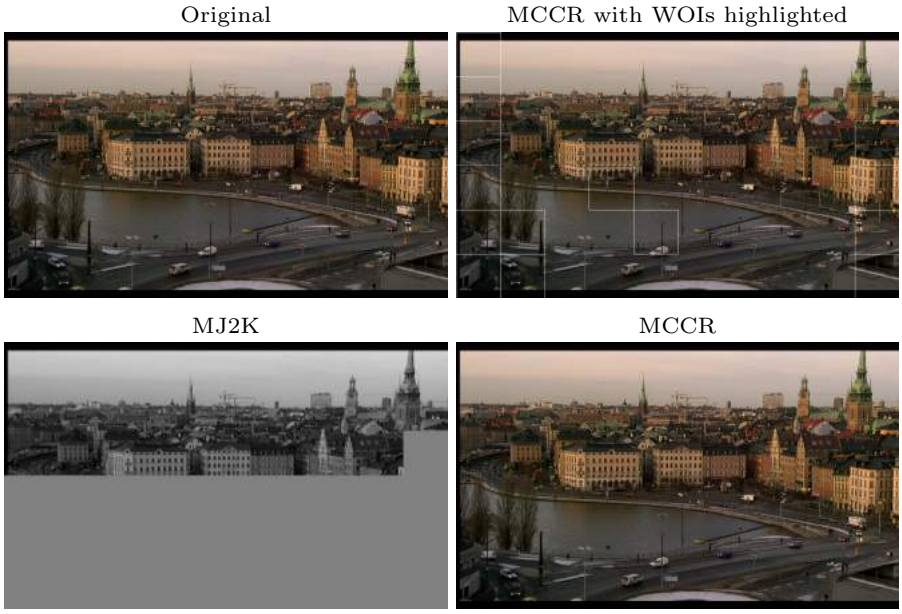



**Fig. 6.** Reconstrucción de la tercera imagen de la secuencia Stockholm utilizando MCCR y Motion JPEG2000 cuando se transmiten 16000 bytes por imagen. En la esquina superior izquierda se muestra la imagen original, en la esquina superior derecha está la reconstrucción obtenida con MCCR, se han resaltado las WOIs que se han actualizado. En la esquina inferior derecha tenemos la misma imagen pero sin resaltar las WOIs y en la esquina inferior izquierda está la reconstrucción obtenida con Motion JPEG2000.

La Figura 6 muestra las reconstrucciones que se han obtenido para MCCR y Motion JPEG2000 cuando se transmiten 16000 bytes por imagen de la secuencia Stockholm. Igual que ocurría en Speedway, la calidad visual que proporciona MCCR es superior a la que ofrece Motion JPEG2000.

## 4 Conclusiones y trabajo futuro

El trabajo que se ha realizado hasta este momento demuestra que MCCR puede llegar a ser una alternativa más eficiente que Motion JPEG2000 cuando se requiere obtener una secuencia de imágenes JPEG2000 con poco movimiento desde un servidor JPIP. Lo que diferencia MCCR de otras propuestas existentes, es que puede ser compatible con cualquier servidor JPIP estándar sin tener que realizar ninguna modificación en el servidor.

Las líneas de investigación relacionadas con MCCR que se plantean como trabajo futuro son las siguientes:

- Mejorar la estimación del movimiento estudiando otros estimadores que puedan mejorar el rendimiento y las reconstrucciones obtenidas actualmente con MCCR.
- Estudiar el comportamiento de MCCR con precintos de menor tamaño. El tamaño de los precintos determina el tamaño de la WOI más pequeña que podemos solicitar al servidor, por lo tanto, en secuencias de imágenes con movimientos complejos, podríamos obtener mejores resultados si el tamaño de los precintos fuese más pequeño.

## References

1. International Organization for Standardization: Information Technology - JPEG 2000 Image Coding System - Part 1: Core Coding System (2004)
2. International Organization for Standardization: Information Technology - JPEG 2000 Image Coding System - Part 9: Interactivity Tools, APIs and Protocols (2005)
3. Müller, D., Nicula, B., Felix, S., Verstringe, F., Bourgoignie, B., Csillaghy, A., Berghmans, D., Jiggens, P., García-Ortiz, J.P., Ireland, J., Zahniy, S., Fleck, B.: JHelioviewer. Time-dependent 3D visualisation of solar and heliospheric data. Astronomy & Astrophysics **606**, A10 (Sep 2017). https://doi.org/10.1051/0004-6361/201730893

# Adaptive Streaming Algorithms and Network Protocols

Teresa Santamarìa-Lòpez

Universidad de Almería, Sacramento S/N, Almería, España
tsantamaria710@gmail.com

**ABSTRACT:** Two Actually the quality and speed of the transmission capacity of users has become one of the problems that will challenge the architecture of server networks and their administrators, so adaptive transmission is proposed as a solution since it allows Dynamic adaptation of the velocity of the flow velocity and we will focus on the type of architecture based on the receiver, that is, an adaptation algorithm that does not delay the information to pass from one layer to another or with server requests. Adaptive transmission in mobile networks is also possible and the main difference of these algorithms is the amount of initial information parameters required for their operation. After this information, a performance experiment was performed where video signals were made through each of the algorithms mentioned above. Three of these types of algorithms presented a passive input, while the adaptation algorithms did not affect the behavior of the output beyond the segment where it began to run.

KEYWORDS**:** Streaming, HTTP, Algorithms, Protocols

## 1 Introduction

A comparison was made between the technologies of the existing adaptation algorithms, among which we find the PANDA Festive [6] and [7] the ABMA + [8] Y, the BBA [9] and BOLA [10]. The BBA [9] and BOLA [10] are based on the Buffer-based adaptation, the PANDA 6 is a performance-based adaptation of four stages. The ABMA algorithm demonstrates the adaptation based on time, within the tests there are 4 types of inputs: Live transmission and stored, within this transmission under normal conditions and in high network traffic conditions. The live and saved transmission does not present important changes when the algorithms are presented, the algorithms of adaptation based on BUFER presented the best performance, that is to say the algorithm BBA and BOLA.

### 1.1 Adaptive Streaming on HTTP

To start to discuss some of the algorithms that exist, we must first analyse the concept of adaptive streaming, then we will see its use through the HTTP protocol. Then we will see implementation of adaptive streaming across networks that typically found in homes and finally a case of comparison of mobile networks. The concept of

streaming has become much more popular years when it's comparit information through video with a lot of user, typically a transmission is made live and in real time, however; There is also the option of storing the transmission to be stored and possibly then request the content many times you want. However, since it is a technology that currently have high demand both quality and speed of transmission capacity of users is one of the issues that will test the architecture of networks of servers and their administrators. Within the large amount of technologies, here we will detail the known as adaptive streaming, it is characterized by allowing that the speed of the stream rate adapts dynamically depending on the changing conditions of the network. At the same time various types of architect of this there are type of adaptive technology, but in this case will be presented the approach of the receiver-driven approach, in which the content files are subdivided into segments, and each of these segments sent to different speeds. In this technology the customer has the task of selecting the rate of speed for each of the segments.

The next challenge that you can appreciate in this technology is that to ensure optimum quality of the steaming should calculate and contain unnecessary fluctuations in data. Finally, to restrict the maximum possible the delay of transmission, independent of the delay of network or equipment is available, must be that the user requests not bounce between different servers as if content makes it along transmission, in this case requests should particularly have a direct channel with the server that provides the content. To solve the various problems mentioned above, there are many algorithms designed and studied around the world that are able to control the fluctuations of the networks where they are implemented. A particular case is an adaptation algorithm receiver-driven for adaptive streaming not delaying information to pass from one layer to another, or with requests for servers. This algorithm was integrated with implementation prototype for a user based on the DASH MPEG streaming client (Dynamic Adaptive Streaming over HTTP) [1]. MPEG DASH is an international standard which uses the already existing traditional HTTP web servers infrastructure and has become very popular in recent years to try to cope with the growing demand for traffic videos [4].

Under the conditions that evaluated the prototype we collected considerably good results under challenging conditions, and remained even stable if the transmission was given to multiple users simultaneously.

Finally the BBA [9] and BOLA [10] are algorithms based on the buffer, that is, on a map of segments that divides the flow of information and is defined by the available values of maximum and minimum quality. In addition, it increases the average quality of the transmission and avoids application conflicts on the server.

### 1.2 Adaptive streaming in mobile networks

We now turn to observe approaches to Adaptive algorithms in mobile networks. Within this class of network technology that has become known as the key solution for the transmission of videos is called searched (HTTP Adaptive Streaming). In mobile networks globally, approximately 60% of the traffic in 2016 has been to transmit video files; also projected that it will continue to grow and increase with the passing of the years [2]. A lot of that traffic is video-on-demand (VoD) streaming through the Protocol already mentioned have [3].

Over the years, have been developed multiple classes of algorithms looking to increasingly improve the Quality of Experience (QoE) [5] users through the adaptation of transmission speeds. The main difference of these algorithms is the amount of initial information parameters that require for their operation, this amount varies from transmission application layers, as well as the size of the buffer of delays.

Some comparisons between technologies of existing adaptation algorithms, which include three main categories will later be evaluated. Firstly, we have the throughput-based algorithms, such as PANDA Festive [6] and [7], which based their decisions or determinations on the captured TCP exits, which to be used need a sufficient number of tests to run and get the required parameters. Secondly, we have the time-based algorithms as ABMA + [8], these are also based on the same principle of a number of tests for the control parameters, but in this case use the estimated times of discharge of each segment of the streaming transmission. Finally, we have the buffer-based algorithms, such as BBA [9] and BOLA [10], that class of algorithms to observe and react to the size of the buffer of delays from customers. [11] considers a practical three kinds of adaptive algorithms already mentioned, debate highlighting the General characteristics of each and explaining more to fund their operating principle.

### 1.3 Throughput-based adaptation

The scheme in which these algorithms are based is based on an Adaptive model of four stages, where the bandwidth of the network availability is estimated using first-order noise filters to prevent errors of estimation in the variations of the output. Then the rate of transmission of the video speed is indicated based on discretized output of the previous point. And in this way following a predictive scheme is expected once the time between request-transmission is calculated.

Within any specific to this type of algorithms variants have the aforementioned PANDA [6], which modifies the model of four stages in two parts. A much more proactive test facility, which is used to reduce as far as possible the fluctuations of rates of change of the speed transmission is used in the estimation stage. The second modification to the conventional paradigm is the step of prediction, where by a controller that uses the maximum size of the buffer client can match the estimated download time with considered solicitud-salida of the server time.

### 1.4 Buffer-based adaptation

Within this category we have the aforementioned BBA and BOLA. BBA is a type of well-known algorithm, which is based on a map of segments that divides the flow of information, said map or drawing is defined by two parameters, which are the available values of maximum and minimum quality available respectively d and the network, and the Strip is expected to keep the size of the segments to be within the parameters of quality control without affecting the speed of transmission.

On the other hand, we have the algorithms of the BOLA class, which is characterized by using a block of optimization of Lyapunov equations to calculate the optimal transmission speed of each segment. This algorithm was made to try to increase the average quality of transmission, avoiding or isolating potential occurrences of conflict of applications on the server.

## 2 Time-based adaptation

For this kind of algorithms, is the main control parameter is the discharge of each segment or packet transmission time. This differentiates them from class Throughput-based algorithms. The aforementioned algorithm ABMA +, is part of this category, and is defined as an algorithm for adaptation and management of buffer, which selects the representation of video based on the probability predicted based on stagnation of the transmitted segments. The algorithm runs continuously estimated download times of the segments in which transmission packets are divided. In addition, it uses a flat pre calculated based on the buffer to select the maximum representation of possible video, which guarantees the quality of the transmission and prevents stagnation of information.

**Table 1:** Characteristics of video signals Characters

| Repre-sentation index | esolution | BBB Max encoding rate (Kbps) | TSA Max encoding rate (Kbps) | RBPS Max encoding rate (Kbps) | CDF 0.05 Quantiles |
|---|---|---|---|---|---|
| 1 | 320x240 | 129 | 128 | 149 | 0.01 |
| 2 | 480x360 | 378 | 330 | 395 | 0.05 |
| 3 | 854x480 | 578 | 754 | 700 | 0.1 |
| 4 | 1280x720 | 985 | 1331 | 1536 | 0.25 |
| 5 | 1280x720 | 1536 | 2048 | 2048 | 0.5 |
| 6 | 1920x1080 | 2353 | 2764 | 2560 | 0.75 |
| 7 | 1920x1080 | 2969 | 3481 | 3072 | 0.95 |

### 2.1 Performance Experiment

The following figure shows the characteristics of video signals used for the experiment.

These signals were made through algorithms of each one of the classes that we saw earlier. An example of the behaviour of three kinds of algorithms below is on a passive input.

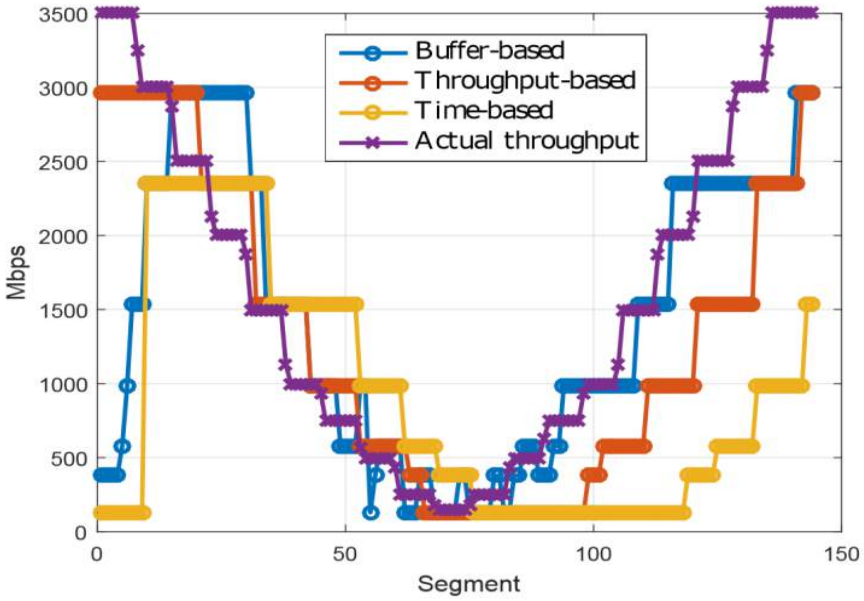



**Figure 1:** Outputs of systems regardless From [11]

You can see that outputs of systems regardless of the class of adaptive algorithm used does not affect the behaviour of output beyond the segment where he began to run.

For tests that were implemented are distinguished four types of entries, live broadcast and saved, and within these two transmission have in normal conditions and in conditions of high network traffic. The first test was carried out to identify the adaptability of the algorithm.

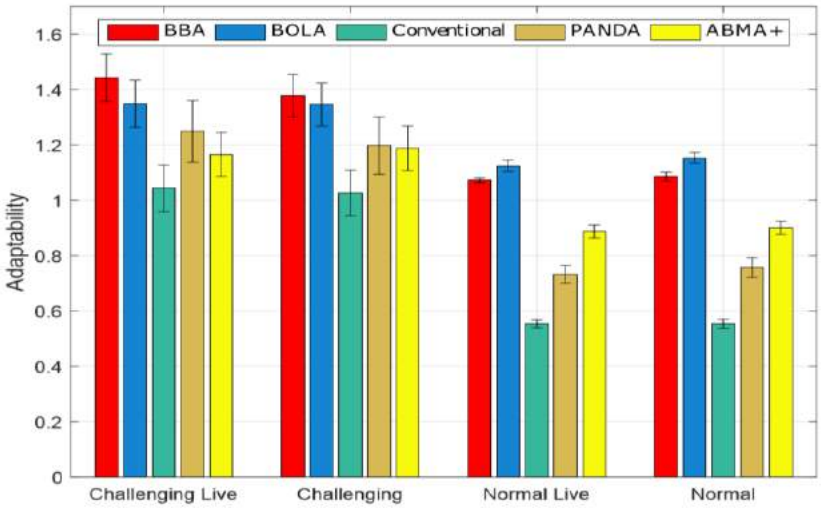



**Figure 2:** Transmission live From [11]

For tests that were implemented are distinguished four types of entries, live broadcast and saved, and within these two transmission have in normal conditions and in conditions of high network traffic. The first test was carried out to identify the adaptability of the algorithm.

It is can demonstrate that the behaviour of a transmission live and saved does not present major changes when the algorithms are implemented. Below are the results for the adaptability of the frequency.

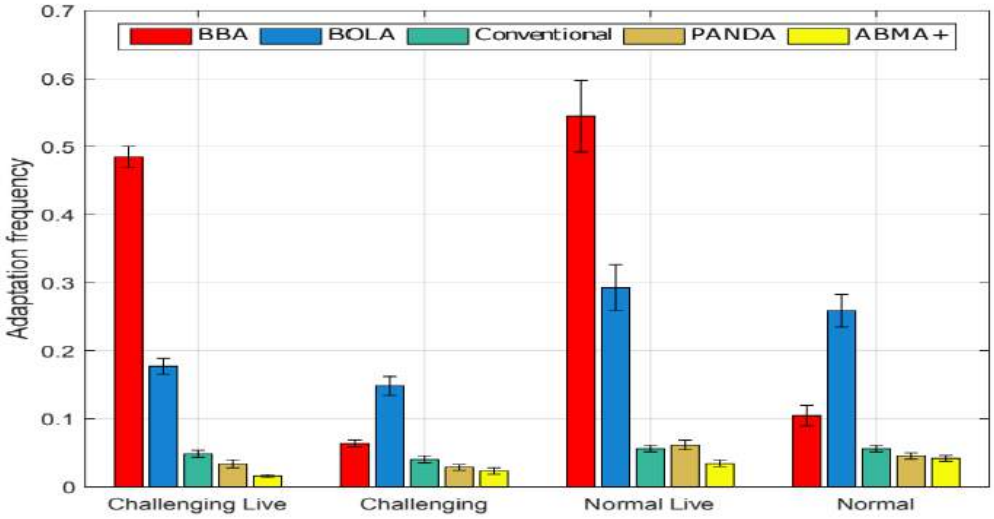



**Figure 3:** Transmission live From [11]

For these parameters clearly class buffer-based adaptation algorithms presented the best performance. Then the adaptation to amplitude.

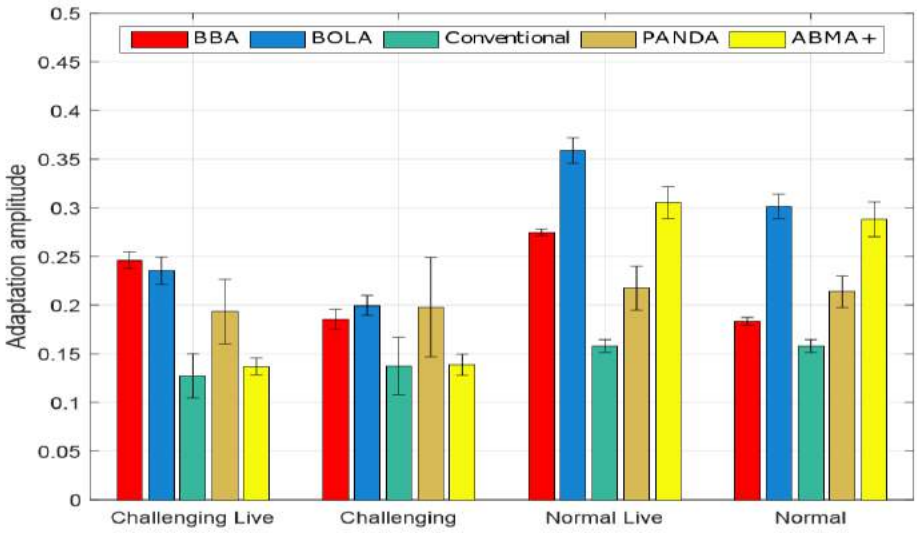



**Figure 4:** Transmission live From [11]

From these findings we can conclude that the best adaptation values are also presented by the BBA and BOLA algorithms.

## 3 Related Work

According to Kathie Nichols and Van Jacobsen, in 2012 they discovered that the crucial point of the high traffic of information sent over the Internet was generated in the queue and there were two critical points: the interval and the threshold. So they created the technique called CODEL, which consists of managing the queue by means of a time tracking, that is, the packets are randomly decremented if an interval value has more delays than arrival points. This method presented a superiority with respect to the general latency specifically in the wireless access links

fcCODEL, was created by Dave Taht, Eric Dumazet, Jim Gettys, with the objective of generating an equal effect in the different flows through the queue, that is, dividing the queue into 1024 subqueues by default and consequently each new flow is randomly delegated to each separate queue and in this CODEL is assigned, which controls TCP congestion

According to a study by the Universidad Javeriana located in Cali, Colombia, a possible solution is to give the opportunity for the routers and switches in the network to behave differently according to the types of service while the traffic passes through

the network. This is known as Differentiated Services (DiffServ) or QoS, which allow coexistence in the network without consuming the bandwidth between the different actors, which means that important flows are used before lower priority flows and that there is also a control of the amount of bandwidth and races between the applications.

Another solution to the problems developed in this article is raised by the School of Telecommunications Engineering of Valencia in 2013 and proposes the introduction of MPLS in the trunk network of access transmission since all types of traffic can be transported between they because they support ATM, IP and TDM links. This is done by means of an emulated circuit indicator header and a system for transferring the clock signal based on RTP (Real-Time Transport Protocol).

## 4 Proposal

In one study, three types of algorithms were analyzed in a passive entry of four types: live transmission and storage, within which two others were carried out: transmission under normal conditions and in high network traffic conditions.

The first test was done with a live transmission and saved, which show no alteration in the implementation of the algorithms.

The second test was carried out to check the adaptability of the frequency in which it was evidenced that the algorithms based on buffer have the best performance, that is, the BBA and BOLA algorithms.

From this analysis, it is obtained as a result that the adaptation algorithm from the segment in which it is started does not affect the behavior of the output.

## 5 Results

The algorithms of adaptation based on class buffer presented the best performance. From these findings, we can conclude that the best adaptation values are also presented by the BBA and BOLA algorithms.

The critical points in the network are presented in the queue, especially in the interval and the threshold of this. For the flow of the types of service (Video, audio, photo, etc.) to increase and not find obstruction during transport the routers and switches must have a differential behavior these. The MPLS allow the transport of all types of traffic, which facilitates the reproduction and consumption of the users of the different types of services

Most of the proposed solutions propose that the best option is based on the division of information into packages and then prioritize it so that transportation does not become obstructed and therefore the reproduction is fluid.

## 6 Conclusiones

Buffet-based algorithms present the main performance parameters of adaptability and stability, especially for small buffers, which are the most common transmission environments.

It is expected that the HTTP protocols, taking the configurations in the receiver, a case study, however, given the algorithms shown below, improve performance if you implement a BBA or BOLA algorithm.

The CODEL technique decreases the packets in the time interval that does not have enough points of arrival in the wireless access links.

fcCODEL controls TCP congestion by delegating each subqueue (subqueals exist because the queue was previously divided into 1024 parts) a new stream.

The QoS admit that the width of the band is not consumed among the different actors of the network and in this way the flows are prioritized so that the important ones are transported first and this allows that there are no obstructions in the flow of the different applications .

MPLS support ATM, IP and TDM links, which means that they transport all types of traffic through an emulated circuit indicator header and a system to move the clock signal based on RTP (Real-Time Transport Protocol).

It can be concluded that the algorithms buffer-based, present the major performance parameters of adaptability and stability especially for small buffers, which are the most common streaming environments.

Protocols of HTTP, taking the settings on receiver-driven, a case of studio, however given the algorithms shown subsequently showed, is expected to improve performance if you implement an algorithm BBA or BOLA

## Referencias


1. Miller K, Quacchio E, Gennari G, Wolisz A. Adaptation algorithm for adaptive streaming over HTTP. 2012 19th International Packet Video Workshop. 2010, pp 173-178.
2. CiscoVisualNetworkingIndex.2017. Global mobile data traffic forecast update, 2016–2021. white paper (Feb.2017).
3. Sandvine.2015. Global Internet Phenomena: Asia-Pacific&Europe. report (Sept. 2015).
4. Sodagar.2011. The MPEG-DASH Standard for Multimedia Streaming Over the Internet. IEEE MultiMedia18,4(April2011),62–67
5. M.Seufert,S.Egger,M.Slanina,T.Zinner,T.Hobfeld,andP.Tran-Gia.2015. A Survey on Quality of Experience of HTTP Adaptive Streaming. IEEE Commun. Surveys Tuts.17,1(2015),469–492.
6. Z.Li,X.Zhu,J.Gahm,R.Pan,H.Hu,A.C.Begen,andD.Oran.2014. Probe and Adapt:Rate Adaptation for HTTP Video Streaming At Scale. IEEE J. Sel. Areas Commun.32(April2014).
7. Bruce P. Douglass. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. DOI: http://dx.doi.org/10.1007/3-540-65193-429
8. JunchenJiang,VyasSekar,andHuiZhang.2012. ImprovingFairness,Efficiency, andStabilityinHTTP-basedAdaptiveVideoStreamingwithFESTIVE.InProc. ACM Int. Conf. on Emerg. Net. Exper. and Techn. (CoNEXT).97–108.
9. Te-Yuan Huang, Ramesh Johari, Nick Mc Keown, Matthew Trunnell, and Mark Watson.2014. A Buffer-based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service. InProc. ACM SIGCOMM.
10. 10]K.Spiteri, R.Urgaonkar,andR.K.Sitaraman.2016. BOLA: Nearoptimalbitrate adaptation for online videos. In IEEE INFOCOM
11. Karagkioules T., Concolato C., Tsilimantos D., Valentin S.,A Comparative Case Study of HTTP Adaptive Streaming Algorithms in Mobile Networks. NOSSDAV. 2017

# Improving the performance of vegetable leaf wetness duration models in greenhouses using decision tree learning

Hui Wang[1]

[1] Centro Mixto CIESOL, University of Almeria
jorgesanchez@ual.es
hw646@ual.es

**Abstract.** Leaf Wetness Duration (LWD) is a key factor when researching greenhouse management due to its sources from over irrigation, rainfall, dewfall. However, LWD estimation is often imprecise because of the lack of a calibrated threshold and the variable decisions made by the researcher. Therefore, this study uses the decision learning tree method (DLT) for four popular LWD models' optimization by acquiring a reasonable threshold from a large dataset and choosing input variables based on calculating their variable importance - RH threshold model (RHM), the dew parameterization model (DPM), the classification and regression tree model (CART) and the neural network model (NNM).

**Keywords:** Air temperature, relative humidity, leaf wetness duration, decision learning tree method.

## 1 Introduction

Research on leaf wetness model emphasized that performance of empirical models varies from place to place, and therefore that they require specific calibration according to the climate of the region [1]. Finding a method for rapidly and effectively calibrating the threshold is important. Data mining algorithms can be used to classify whether a leaf is wet or dry. The classifying value is the same as the calibrated threshold.

Decision tree learning is one of the most successful learning algorithms due to its various attractive features: simplicity, comprehensibility, no parameters, as well as being able to handle mixed-type data [2]. A class label and a tuple of attributes are used to train labeled data. Its advantage is its vast research space and its recursive processing until all instances in a subset belong to the same class. It not only uses to the leaf wetness model has single variable, but also suitable for the model has multi variables or categories. The calibrated threshold of leaf wetness model get from the tree node by classifying the objective variable using a class label (wet or dry) and a tuple of attributes (model variable names).

Determining the most important variables in a model—which to include (and exclude) in the model, and which of the included variables contribute the most to

prediction—is critical from both the practical and theoretical perspectives [3]. For NNM is a network structure including the elements-input, output and intermediate layers. The number of intermediate layers, how many elements are contained therein, and the connection between elements are user-selectable; this gives the network flexibility in pattern recognition during model development, but this leads to alterable results. Francl and Panigrahi [4] analyzed the sensitivity of artificial neural network models of wetness status at the wheat flag leaf level to individual input variables, therein estimated leaf wetness duration and relative humidity were very important factors, as expected, and temperature, solar radiation, and time of day also were influential. Of lesser importance were wind speed and precipitation, but predictions were always correct when it rained. Wind direction apparently contributed little to model accuracy because values were close to the 50% level achievable by random chance. However, Stella et al [5] proposed the inputs of artificial Neural Network (ANN) for leaf wetness estimation need to be the mostly correlated climate variables with leaf wetness- air temperature, relative humidity, rainfall, wind speed, solar radiation. Therefore, finding a method for rapidly and effectively calibrating the leaf wetness model 'threshold and justifying the importance and sensitivity of the inputs of NNM is necessary. Decision tree learning supplies a function of estimates predictor importance for the tree by totalling the changes in the mean squared error (MSE) caused by the splits for each predictor and then dividing the total by the number of branch nodes. The variable importance associated with this split is computed as the difference between MSE for the parent node and the total MSE for the two child nodes.

The objective of this thesis would contribute on constructing a method to have a local threshold of leaf wetness model and compare their accuracy before and after calibration using reference data obtained in dedicated experimental trials (Figure 1).

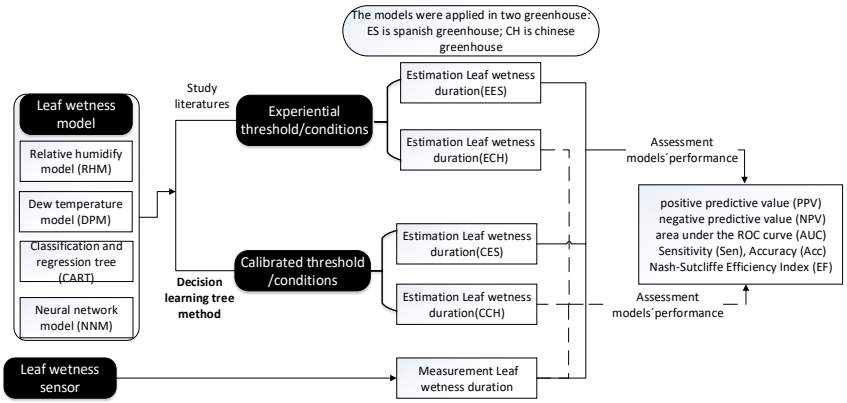



Figure 1 The workflow of evaluating the calibrated thresholds effective on LWD models' performance

## 2 Objective

The objectives of the present thesis can be divided into two parts. On the one hand, simulation leaf wetness duration using two weather station data (Beijing, China and Almeria, Spain), as well as calibrating the thresholds and input parameters of the models by decision tree learning. On the other hand, comparing the results of different leaf wetness models aims to choose the one has better performance.

Firstly, four leaf wetness models were researched. Thus, we have to finish some work are descripted as below.

1. Study and research four leaf wetness models: relative humidity model (RHM), dew temperature model (DPM), classification and regression tree model (CART) and neural network model (NNM).

2. Simulation of leaf wetness duration based on the weather data and experience threshold and parameters is to compare the results from the models with calibrated conditions.

Next step is analysing the results to prove the calibrated condition is useful to improve the models' performance using statistical indicators. The objectives we have to do:

1. Choose the statistical indicators could explain the models' performance.

2. Analysing the reason of the different results of the models before and after calibration.

3. Evaluation results of the models after calibration using new data set.

## 3 Research progress

This part introduces how is going for the objectives. For the first objective, some work we had done as below:

1. *Simulation leaf wetness models by the four leaf wetness models with experience threshold and parameters.*

   We used the data set from two weather stations in different location simulate the leaf wetness duration during three plant seasons (in the Spanish greenhouse from April 2016 to December 2016 and in the Chinese greenhouse from April 2014 to November 2014). The models were executed by Matlab 2018a. A decision tree algorithm works by splitting a data set in order to train

a model through a recursive partitioning process, and then the model is used to predict the value of a target variable based on the independent variables (Figure 2).

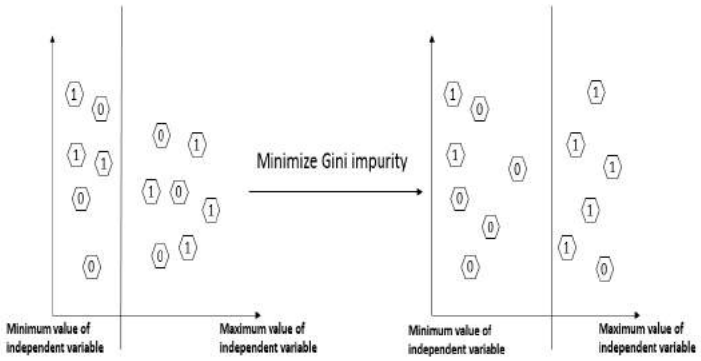



**Figure 2.** Process of optimizing threshold of leaf wetness model by decision learning tree

2. *Comparison of the results from the four models after calibration by six statistical indicators*

Table 1 Performance statistic values of the LWD models in two greenhouses: relative humidity model (RHM), dew temperature model (DPM) and neural network model (NNM)

| Location | Spain | | | | China | | | |
|---|---|---|---|---|---|---|---|---|
| Index/Mode | RHM | DPM | CART | NNM | RHM | DPM | CART | NNM |
| PPV | 0.71 | 0.71 | 0.77 | 0.82 | 0.85 | 0.86 | 0.84 | 0.90 |
| NPV | 0.72 | 0.71 | 0.76 | 0.87 | 0.80 | 0.78 | 0.87 | 0.92 |
| Sen | 0.62 | 0.64 | 0.68 | 0.85 | 0.75 | 0.80 | 0.81 | 0.89 |
| Acc | 0.73 | 0.73 | 0.76 | 0.88 | 0.87 | 0.86 | 0.86 | 0.92 |
| AUC | 0.73 | 0.74 | 0.77 | 0.90 | 0.79 | 0.83 | 0.84 | 0.94 |
| EF | 0.83 | 0.83 | 0.85 | 0.92 | 0.85 | 0.86 | 0.88 | 0.94 |
| MAE | 5.45 | 5.46 | 3.46 | 1.85 | 3.42 | 3.29 | 2.75 | 1.30 |
| SD | 6.11 | 6.11 | 4.11 | 2.25 | 4.43 | 4.20 | 3.51 | 1.62 |

Note: performance statistics indexes include: positive predictive values (PPV), negative predictive values (NPV), sensitivity (Sen), Accuracy (Acc), the area under Receiver Operating Characteristic curve (AUC), EF (effective fitting), mean absolute error (MAE) and Root mean square error (RMSE).

The data were acquired from March 2017 to February 2018 in Spain and from December 2014 to October 2015 in China - these were used to assess the calibration models. Depending on the evaluation criteria, there was a top-down performance ranking for the leaf wetness models (Table 1). Of the LWD models in Spanish greenhouse, NNM was the most efficient at estimating LWD, with PPV of 0.82 and NPV of 0.87, while also smallest error between estimated LWD and measured LWD existed in NNM's results, is 1.85h (MAE) and 2.25h (SD). Furthermore, the evaluation results by the other two LWD models are similar, with PPV of 0.71 and 0.71, while also, with NPV of 0.71 and 0.72. In the Chinese greenhouse, the NNM gave an excellent evaluation result, with PPV of 0.90, NPV of 0.92, MAE of 1.30h and SD of 1.62h.

## 4 Conclusion

In the work is intending to contribute on in the greenhouse of simulation leaf wetness duration, with the purpose of improving the disease model's effectiveness as it is an important factor leading to some fungi disease happen. So far, the contribution has been done that calibration of four leaf wetness models, comparison of the models' performance, assessment of the importance for specific calibration according to the climate of the region.

## References

1. Sentelhas, P. C.; Dalla Marta, A.; Orlandini, S.; Santos, E. A.; Gillespie, T. J.; Gleason, M. L. Suitability of relative humidity as an estimator of leaf wetness duration. *Agric. For. Meteorol.* **2008**, *148*, 392–400, doi:10.1016/j.agrformet.2007.09.011.
2. Su, J.; Zhang, H. A Fast Decision Tree Learning Algorithm. In *21st national conference on Artificial intelligence*; 2006; Vol. 1, pp. 500–505.
3. Braun, M. T.; Oswald, F. L. Exploratory regression analysis: A tool for selecting models and determining predictor importance. *Behav. Res. Methods* **2011**, *43*, 331–339, doi:10.3758/s13428-010-0046-8.
4. Francl, L. J.; Panigrahi, S. Artificial neural network models of wheat leaf wetness. *Agric. For. Meteorol.* **1997**, *88*, 57–65, doi:10.1016/S0168-1923(97)00051-8.
5. Stella, A.; Caliendo, G.; Melgani, F.; Goller, R.; Barazzuol, M.; La Porta, N. Leaf Wetness Evaluation Using Artificial Neural Network for Improving Apple Scab Fight. *Environments* **2017**, *4*, 42, doi:10.3390/environments4020042.