Modelling and Estimation of Environmental and Energy Standards in Buildings

Ricardo Filipe Pacheco de Oliveira, Instituto Superior Técnico

Abstract

The efficient use of energy resources is now a key imperative for the sustainability of the economy, the environment and the current level of comfort and quality of life. The development of energy management systems that lead to significant energy savings opens a huge range of opportunities in the field of information technologies. The maintenance of these parameters of comfort is mainly related to building acclimatization, which means that it's intrinsically linked to the HVAC equipment. The evaluation of this behaviour through the use of tools to study the thermal performance of buildings, require nonlinear time-consuming analyzes in terms of implementation in terms of computational weight. As a practical alternative, we intend to use models created from long time series of environmental variables and energy, obtained from the processed data from the existing technical and energy management systems, which are representative of buildings nationwide. Based on these models the intention is to develop algorithms that allow firstly an efficient encoding of the vast collection of signals (time series) and then the estimation of missing data through the reconstruction based on history. Ultimately the objective is to predict the thermal behaviour and energy requirements related to the control of thermal loads. The algorithms should be simple to apply, preferably linear in nature, and with low computational weight, to allow its implementation in "real time" and its integration in the existing building technical and energy management systems.

Index Terms - energy efficiency in buildings, technical and energy management, time series analysis, prediction of the thermal behaviour of the building and associated energy consumption.

I. INTRODUCTION

The efficient use of energy resources is an imperative for the sustainability if the actual indexes of comfort and quality of life. Observing energy consumption and environment parameters long time series (temperature, air particles, etc.), we intend to develop algorithms for:

a) Efficient codification of the vast collection of signals obtained through time in a large facility;

b) Estimation of missing data through reconstruction exploring data redundancy, in particular its low dimensionality; use of this methodology to predict energy consumption;

c) Unsupervised aggregation of data into classes that show regular and spurious behaviour of the facility;

Within this context this paper is divided into 4 additional chapters; on chapter 2 we characterize the building, the energy management equipments and the observations, on chapter 3 we develop a method to predict the energy consumption of the HVAC (air conditioning) using the existing data of the HVAC, on chapter 4 we apply 3 methodologies to estimate the missing data and on chapter 5 we use a clustering algorithm to determine possible relations between variables.

II. SYSTEM DESCRIPTION

The building in question is a retail surface within the urban Lisbon area, with a warehouse and an outside parking, with a constructed area of 1.300m², of which 1.000m² correspond to the sales area.

2.1 Energy Management

The Technical and Energy Management System (TEMS), developed and installed by the author, manages autonomously all the equipments existing on this facility and also reads all the data that will be used in the construction of the mentioned models. The system hardware uses industrial equipment for a question of reliability, warranty and scalability.

![TEMS Architecture](SOURCE: TEMS)

2.2 Data Validation

An essential step is to confirm the data is in fact reliable. For this purpose we compare the energy consumption values from the electrical energy supplier invoices and the data gathered by the TEMS, between 10-08-2010 and 25-01-2012. This comparison has an average error of 5,4%, easily explained by the different taxation periods. Extending the comparison period to 12 months, the error decreases to 0,78%, so we conclude the observations are valid.
2.3 Characterization of the Observations / Variables

Having validated the readings, we now consider the most recent period available between 01-01-2013 and 28-02-2014.

![Fig. 2](image-url) – Energy consumptions for the period between 01-01-2013 and 28-02-2014

1. total energy
2. energy from the negative freezers
3. energy from the positive freezers
4. HVAC energy
5. energy from the carton compact machine
6. energy from the interior lights
7. energy from the exterior lights
8. energy from the bread oven
9. energy from the negative freezer from the bread
10. energy from the fish shop
11. energy from the butcher [SOURCE: TEMS]

### Table 1 – Energy consumption per month and equipment for 2013 [SOURCE: TEMS]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>96,371</td>
<td>442</td>
<td>11,46</td>
<td>4,127</td>
<td>35,25</td>
<td>474</td>
<td>2,230</td>
</tr>
<tr>
<td>MINIMUM</td>
<td>1,010</td>
<td>400</td>
<td>2,000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MAXIMUM</td>
<td>96,000</td>
<td>336</td>
<td>22,36</td>
<td>27,68</td>
<td>2,090</td>
<td>24,70</td>
<td>6,38</td>
</tr>
<tr>
<td>STANDARD DEVIATION</td>
<td>28,076</td>
<td>500</td>
<td>6,622</td>
<td>6,74</td>
<td>321</td>
<td>8,005</td>
<td>1,605</td>
</tr>
<tr>
<td>VARIANCE</td>
<td>427,91</td>
<td>125</td>
<td>39,55</td>
<td>22,52</td>
<td>105</td>
<td>76,68</td>
<td>2,574</td>
</tr>
</tbody>
</table>

A brief analysis of the time series allows us to conclude:

- **Positive and Negative Freezers, and HVAC**: average increase of 30% from winter to summer; during the day the consumption is higher within the opening hours; high dependency on weather conditions;
- **Interior Lights**: periodic variable from day to day, since operation schedule is the same;
- **Exterior Lights**: although a periodic variable between days, there is an average increase of 300% from summer to winter, due to the shorter period of sun exposure;
- **Carton Press Machine**: it’s an equipment of random and sporadic use, with no observable periodicity;
- **Bread Oven**: it’s an equipment where only 50% of the time there is a determined schedule;
- **Fish Shop and Butcher**: continuous and relatively stable consumption between the day;
- **Total**: Increase of 6.3% from summer to winter. Concerning the HVAC, we proceed with an identical analysis.

![Fig. 3](image-url) – Temperature for the period between 01-01-2013 and 28-02-2014

1. interior temperature
2. exterior temperature
3. supply air temperature
4. return air temperature [SOURCE: TEMS]

### Table 2 – HVAC data related to 2013 [SOURCE: TEMS]

<table>
<thead>
<tr>
<th>2013</th>
<th>Temperature (ºC)</th>
<th>Temperature (ºC)</th>
<th>Supply Air Temp (ºC)</th>
<th>Return Air Temp (ºC)</th>
<th>Access Rate</th>
<th>HVAC Cost (€/GWh)</th>
<th>HVAC Cost (€/GWh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>20,026</td>
<td>16,116</td>
<td>19,88</td>
<td>19,28</td>
<td>40,20</td>
<td>78,95</td>
<td>625,69</td>
</tr>
<tr>
<td>MINIMUM</td>
<td>12,30</td>
<td>2,20</td>
<td>7,80</td>
<td>7,20</td>
<td>12,60</td>
<td>14,50</td>
<td>400,00</td>
</tr>
<tr>
<td>MAXIMUM</td>
<td>27,30</td>
<td>58,85</td>
<td>40,00</td>
<td>29,50</td>
<td>59,80</td>
<td>100,00</td>
<td>1,120,00</td>
</tr>
<tr>
<td>STANDARD DEVIATION</td>
<td>2,05</td>
<td>5,84</td>
<td>2,53</td>
<td>3,20</td>
<td>5,23</td>
<td>5,84</td>
<td>9,60</td>
</tr>
<tr>
<td>VARIANCE</td>
<td>4,20</td>
<td>34,14</td>
<td>27,39</td>
<td>10,22</td>
<td>23,26</td>
<td>37,84</td>
<td>10,515,13</td>
</tr>
</tbody>
</table>

The most relevant remarks are:

- **Interior Temperature**: corresponds to the set-point defined by the user and is relatively stable across the year;
- **Exterior Temperature**: depends on the weather conditions, so it varies greatly from winter to summer, but has some periodicity from day to day within the season;
- **Supply Air Temperature**: corresponds to the temperature of the air supplied to the interior of the facility and depends on the set-point, current interior temperature, exterior temperature and the percentage of fresh air; it’s the least deterministic variable of all;
- **Return Air Temperature**: corresponds to the temperature of the air retrieved from the interior of the facility and it’s very similar to the interior temperature;
- **Interior Humidity**: relatively constant through all the day and all the year;
- **Exterior Humidity**: depends strongly on the weather conditions, but retains some periodicity from day to day;
- **Air Quality**: depends almost exclusively on the amount of persons inside the facility;
- **Percentage of Fresh Air**: also depends greatly on the amount of persons inside the facility and the need to introduce fresh air;

We conclude that most variables are periodic from day to day and suffer from seasonal effect.

2.4 HVAC Potential of economy

The level of complexity of the price structure of the energy sector in Portugal is very high, depending on:

- Access rate, which includes the fixed term set by the regulator, the power in peak hours, the contracted power and the fixed term for the active energy in each period;
- Active Energy cost in each period;
- Cost of reactive power;

We emphasise the peak period, where in addition to the higher costs per kWh, there is an additional taxation based on the peak power.

It is important to detail the operation of the air conditioning system, since it’s the target of energy optimization. Generally speaking, the acclimatization of the interior space is performed by the supply of heated or cooled air; the return air is in turn collected and returns to the air-conditioning equipment where it is reheated or re-cooled, together with the new air from the exterior.

Generating example charts with the daily progression of the distinct HVAC variables in rigorous summer and winter days we obtain:
The charts make it clear that the energy consumption is high during peak hours, from 09:15 to 12:15 in the summer and from 9:30 to 12:00 and 18:30 to 21:00 in the winter.

Considering the costs at the time, in 2011, consumption during peak hour in the summer was 3.480kWh and in the winter 11.226kWh, which corresponds to a cost of 1.820,60€ with energy and of 2.530,68€ with power, which totals 4.351,29€. Transferring these costs to the full period, for winter 11.226kWh, which corresponds to a cost of 1.298,54€, example by turning the equipment off automatically before and after the peak hour, we would have a cost of only 1.298,54€, which corresponds to savings about 70% for each facility or 305.272€ for 100 facilities.

2.5 Relations between observations

In order to try and establish a relationship between the data we construct the table below with the cross-correlations between all variables:

### Table 3 - Cross correlation table between all HVAC observations

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Strong</th>
<th>Average</th>
<th>Weak</th>
<th>Very weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strong</td>
<td>1.00</td>
<td>0.89</td>
<td>0.39</td>
<td>0.19</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.89</td>
<td>0.40</td>
<td>0.20</td>
</tr>
<tr>
<td>Weak</td>
<td></td>
<td></td>
<td>0.39</td>
<td>0.19</td>
</tr>
<tr>
<td>Very weak</td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
</tbody>
</table>

We conclude the following:

- Strong correlation between temperatures;
- Strong correlation between the total energy and the energy from positive temperature freezers and negative freezers;
- Almost all the variables have moderate or strong correlation, suggesting low dimensionality;

2.6 Construction of the data matrix

As observed, there is some periodicity in the data from day to day, so we'll take a data model in the form of matrix called \( C_b \), where we group each day of each variable in the first 24 lines (19x24 lines); each column corresponds to one day.

### Fig. 6 – Observations for the period between 01-01-2013 and 28-02-2014

Upper left: 1 – total energy / 2 – energy of the negative temperature freezer / 3 - energy of the positive temperature freezer / 4 – HVAC energy / 5 – compact machine energy / 6 – interior lights energy / 7 - exterior lights energy / 8 – bread oven energy / 9 - energy of the bread negative temperature freezer / 10 – fish shop energy / 11 – butcher energy

Upper right: 1 – interior temperature / 2 –exterior temperature / 3 – air supply temperature / 4 – return air temperature

Lower left: 1 – interior humidity / 2 – exterior humidity

Lower right: air quality

III. HVAC Energy prediction

In this chapter we intend to validate the hypothesis of predicting the HVAC energy consumption, using the remaining HVAC observations.

3.1 State of the art

There are several simulation tools that enable the study of the thermal behaviour of a building as the "Energy Plus", the "HAP" or the "TRACE", but too complex to be usable in real time. It is in this context that we want to implement a simple methodology, possible to integrate in the TEMS, in order to:

- Estimate the actual energy consumption and thus eliminate the dependency of an external meter;
- Forecast the energy consumption of a given hour or day, when a single or multiple variables are modified (temperature, humidity, etc.) and thus optimize the set point / interior temperature that minimizes the energy;

The real interest is the prediction of the HVAC energy consumption is to create a "data driven" model that allows its energy optimization taking into account the known variables. Looking in more detail the available variables, we have:

- Interior temperature: known; defined by user;
- Exterior temperature: is possible with some rigor to know your hourly evolution throughout the day using the data of the weather forecast;
- Air supply temperature: it's the temperature of air to acclimate the space, which naturally depends on the interior and exterior temperatures; is not known;
Return air temperature: very similar to the interior temperature, so we consider it known;
• Interior humidity: with this type of air conditioning and for the same geographic region, takes an almost constant value, so we consider it known;
• Exterior humidity: we can use the data from the weather forecast to obtain the hourly values;
• Air quality: depends on the number of people within the space, so it is not known;
• Percentage of fresh air: the greater the number of people within the space, the greater the need for fresh air, to overcome the higher levels of CO₂; not known;
• HVAC energy: variable we intend to estimate;
We will focus on the energy, assuming that the values of the remaining variables are known, and we leave for another analysis the problem of their determination.

3.2 Efficient data codification
To calculate or predict the HVAC energy we privilege linear models for various reasons:
• simplicity of implementation and execution, since the method is expected to run in real time by the TEMS;
• We don’t have a model to characterize this system;
• The redundancy of the observation, since we assume the data is correlated, or in other words, it’s low rank;

An important question that arises is the actual numeric value of data and its different scales. To avoid this problem, it is chosen to use the standardized data, considering its mean and standard deviation:

\[ c'_i = \frac{c_i - \mu(C)}{\sigma(C)}, \]  

where \( c'_i \) is the normalized value, \( c_i \) the original value, \( \mu(C) \) the average of \( c_i \) and \( \sigma(C) \) its standard deviation.

Coming back to the matter of the periodicity of the variables, it is to assume that the data is of low rank, that is, with a small set observations we can represent the full set.

In order to prove this assumption, consider the following:
1 – Decomposition of the observation matrix into singular values \( Wa = U \ast D \ast V' \), where \( U \) is an unitary matrix (\( U \ast U' = I \)) , whose columns form the singular vectors in the left, \( D \) is a diagonal matrix with the singular values of \( Wa \) and \( V \) is also a unitary matrix, whose columns form the singular vectors on the right. \( U \) and \( V \) are orthonormal bases, which means its columns form an orthonormal set of vectors.
2 – Rank reduction to rank “n”, considering only the “n” higher singular values.
3 – Reconstruction of the matrix \( Wa_{rank} = U \ast D_{rank} \ast V' \)

To measure the reconstruction error we consider the Frobenius norm:

\[ \| En_{real} - En_{estimado} \| = \left( \sum_{i=1}^{m} \sum_{j=1}^{n} |e_{ij}|^2 \right)^{1/2}, \]  

Observing the graphic development of the singular values and the corresponding reconstruction error of the matrix, we observe that around rank 5 there is higher fall in the reconstruction error, consistent with low rank data.

3.3 Description of the method
With the previous assumptions, we consider \( Wa \) as:

\[ Wa = \begin{bmatrix} Ena & \ldots \ & Enb \end{bmatrix}, \quad \text{or its alternate representation} \quad Wa = \begin{bmatrix} Wa_1 \ldots \ & Wa_{n} \end{bmatrix} \]  

Considering its single value decomposition

\[ Wa = U \ast D \ast V', \]  

And since \( U \) is in fact a base for the columns of \( W_a \) and of \( W_b \) we can assume

\[ Wb = U \ast Cb \]  

where \( U \) is a base for the columns and \( Cb \) will be simply called coefficient.

Representing the above equation in the form:

\[ \frac{En}{Wb_1} = \begin{bmatrix} En_1 \ldots \ & En_n \end{bmatrix} \ast Cb. \]  

Solving this system in order to \( Cb \), with the known data \( Wb_1 \) and \( Ub \), the solution is provided by the pseudo-inverse:

\[ Cb = (Ub' \ast Ub)^{-1} \ast Ub' \ast Wb_1, \]  

and

\[ En = UEn \ast Cb, \]  

Taking into account our initial assumption that \( Wa \) is low rank, the above equations are rewritten as:

\[ \begin{bmatrix} En(24x1) \\ Wb(12x2016) \end{bmatrix} = \begin{bmatrix} En_1 \ldots \ & En_{2016} \\ Ub_1 \ldots \ & Ub_{2016} \end{bmatrix} \ast Cb(2016), \]  

\[ Cb(2016) = (Ub'(12x2016) \ast Ub(12x2016))^{-1} \ast Ub'(12x2016) \ast Wb(12x2016), \]  

and

\[ En(24x1) = UEn(24x2016) \ast Cb(2016), \]  

where \( rk \) is the required rank

3.4 Results
As already indicated, rank 5 provides the best global estimation, so taking this into account, we estimate the first 15 days of 2014 using the data from 2013.
The time series related to the observed data (real) is shown in blue and the time series of the estimate obtained is shown in red.

Figure 8 – Comparison between the real value and the estimate of the HVAC energy for 6 days.

The following table presents the forecast errors for each of the days, with emphasis on the relative error.

<table>
<thead>
<tr>
<th>Day</th>
<th>Error (absolute mean)</th>
<th>Mean absolute error</th>
<th>Relative mean error</th>
<th>Maximum absolute error</th>
</tr>
</thead>
<tbody>
<tr>
<td>02-01-2015</td>
<td>1.465796</td>
<td>0.1887458</td>
<td>19.7%</td>
<td>0.9993732</td>
</tr>
<tr>
<td>03-01-2015</td>
<td>0.966183</td>
<td>0.1652060</td>
<td>10.0%</td>
<td>0.5408202</td>
</tr>
<tr>
<td>04-01-2015</td>
<td>2.380555</td>
<td>0.4305249</td>
<td>15.7%</td>
<td>0.8460430</td>
</tr>
<tr>
<td>05-01-2015</td>
<td>1.561687</td>
<td>0.2439268</td>
<td>30.8%</td>
<td>0.7824829</td>
</tr>
<tr>
<td>06-01-2015</td>
<td>1.110331</td>
<td>0.1694003</td>
<td>14.4%</td>
<td>0.5998934</td>
</tr>
<tr>
<td>07-01-2015</td>
<td>0.949423</td>
<td>0.1428561</td>
<td>14.0%</td>
<td>0.5367428</td>
</tr>
<tr>
<td>08-01-2015</td>
<td>1.089961</td>
<td>0.1756309</td>
<td>12.1%</td>
<td>0.5701031</td>
</tr>
<tr>
<td>09-01-2015</td>
<td>1.209290</td>
<td>0.1866882</td>
<td>19.4%</td>
<td>0.5308762</td>
</tr>
<tr>
<td>10-01-2015</td>
<td>1.329725</td>
<td>0.2367135</td>
<td>13.5%</td>
<td>0.6612572</td>
</tr>
<tr>
<td>11-01-2015</td>
<td>2.050381</td>
<td>0.3491911</td>
<td>37.3%</td>
<td>0.8497625</td>
</tr>
<tr>
<td>12-01-2015</td>
<td>1.315697</td>
<td>0.1737012</td>
<td>18.7%</td>
<td>0.7820880</td>
</tr>
<tr>
<td>13-01-2015</td>
<td>1.112822</td>
<td>0.1672919</td>
<td>12.6%</td>
<td>0.4524277</td>
</tr>
<tr>
<td>14-01-2015</td>
<td>1.260608</td>
<td>0.1957545</td>
<td>25.0%</td>
<td>0.6187914</td>
</tr>
<tr>
<td>15-01-2015</td>
<td>1.853829</td>
<td>0.2936443</td>
<td>25.7%</td>
<td>0.7837136</td>
</tr>
<tr>
<td>16-01-2015</td>
<td>1.312520</td>
<td>0.1808362</td>
<td>11.1%</td>
<td>0.6820710</td>
</tr>
</tbody>
</table>

Table 4 – Table with the prediction error of the HVAC energy on 15 consecutive days.

In a more detailed analysis of the first 50 singular values of the same 15 days of 2014, we can observe different minimum for each day, although there is some prevalence around the rank 5, as we had assumed.

Fig. 9 – Evolution of the error with the rank for 3 consecutive days.

Although the relative error values around 20% are not ideal, the representation of the time series is relatively accurate, so we considered this model appropriate for the proposed objectives and accept the results as valid.

3.5 Alternate calculation method

In order to see how far we can take this method, we make the simultaneous prediction for ‘n’, using the data available; the equations take the form:

\[ C_b(x \times n) = (U_b^{(rk \times 2)} \ast U_b^{(r \times 2)})^{-1} \ast U_b^{(rk \times 2)} \ast Wb1^{(2 \times 2)} \],

and

\[ E_n(24 \times n) = U E_n(24 \times r \times k) \ast C_b^{(rk \times n)} \],

being rk the same rank 5 already considered.

In the following table we have synthesized the values of the forecast error for the same 15 days.

<table>
<thead>
<tr>
<th>Day</th>
<th>Error (absolute mean)</th>
<th>Mean absolute error</th>
<th>Relative mean error</th>
<th>Maximum absolute error</th>
</tr>
</thead>
<tbody>
<tr>
<td>02-01-2015</td>
<td>1.590352</td>
<td>0.1884486</td>
<td>18.2%</td>
<td>1.1057700</td>
</tr>
<tr>
<td>03-01-2015</td>
<td>0.949552</td>
<td>0.1682568</td>
<td>13.5%</td>
<td>0.4675315</td>
</tr>
<tr>
<td>04-01-2015</td>
<td>2.340749</td>
<td>0.4231998</td>
<td>14.3%</td>
<td>0.8142103</td>
</tr>
<tr>
<td>05-01-2015</td>
<td>1.244055</td>
<td>0.1832568</td>
<td>28.8%</td>
<td>0.6731818</td>
</tr>
<tr>
<td>06-01-2015</td>
<td>1.032173</td>
<td>0.1530912</td>
<td>12.3%</td>
<td>0.5327603</td>
</tr>
<tr>
<td>07-01-2015</td>
<td>0.798632</td>
<td>0.1229798</td>
<td>10.1%</td>
<td>0.4681081</td>
</tr>
<tr>
<td>08-01-2015</td>
<td>1.010727</td>
<td>0.1589560</td>
<td>36.6%</td>
<td>0.5287155</td>
</tr>
<tr>
<td>09-01-2015</td>
<td>1.422802</td>
<td>0.2141164</td>
<td>21.7%</td>
<td>0.5750645</td>
</tr>
<tr>
<td>10-01-2015</td>
<td>1.246698</td>
<td>0.2373303</td>
<td>11.8%</td>
<td>0.5017739</td>
</tr>
<tr>
<td>11-01-2015</td>
<td>1.693088</td>
<td>0.3029517</td>
<td>35.0%</td>
<td>0.6398405</td>
</tr>
<tr>
<td>12-01-2015</td>
<td>1.054836</td>
<td>0.1351245</td>
<td>15.9%</td>
<td>0.6669103</td>
</tr>
<tr>
<td>13-01-2015</td>
<td>0.935186</td>
<td>0.1447224</td>
<td>10.7%</td>
<td>0.4334764</td>
</tr>
<tr>
<td>14-01-2015</td>
<td>0.928508</td>
<td>0.1384058</td>
<td>21.5%</td>
<td>0.4649890</td>
</tr>
<tr>
<td>15-01-2015</td>
<td>1.414956</td>
<td>0.2335919</td>
<td>22.8%</td>
<td>0.5815814</td>
</tr>
<tr>
<td>16-01-2015</td>
<td>1.160832</td>
<td>0.1666970</td>
<td>18.2%</td>
<td>0.5717006</td>
</tr>
</tbody>
</table>

Table 5 – Table with the prediction error of the HVAC energy on 15 consecutive days, estimated simultaneously.

This method also provides acceptable results, although slightly worse than the previous single day prediction.

IV. GLOBAL ESTIMATION OF MISSING DATA

In this new chapter we intended to solve a different problem; to globally estimate the missing data in the time series, which is normally randomly dispersed, typically due to miscommunication or malfunctions.

4.1 Assumptions

In this chapter we consider the complete set of observations, i.e., all values of energy and HVAC in the form of the \( C_b \) matrix already described. As in the previous chapter, we assume that the rank of the matrix is low and therefore we first analyze the results of the single value decomposition and the evolution of the reconstruction error:

![Fig. 10 – Evolution of the singular values of \( C_b \) (left) and evolution of the reconstruction error of the matrix \( C_b \) with the rank (right).]

The ratio between the singular value corresponding to rank 60 and rank 1 exceeds 200, which validates the initial assumption of low rank data.

4.2 Estimation of missing data

We study 3 different methodologies:
The first method is implemented fully, while for the remaining we use the code of authors.

4.4 A - Expectation Maximization algorithm

The first algorithm, developed in the paper "Estimation of Rank Deficient Matrices from Partial Observations: Two-Step Interactive Algorithms", is a method that estimates iteratively the missing data and afterwards the low rank approximation; the process is iterative until convergence.

Consider the observations \( W \) of a low rank matrix \( \hat{W} \), of \( M \times N \) such as rank \( R < \min(M,N) \), corrupted with white Gaussian noise. The maximum likelihood estimate \( \hat{W} \) of \( W \) is:

\[
\hat{W} = \arg \min_{\hat{W}} \| W - \hat{W} \|_F, \tag{14}
\]

where \( \| \cdot \|_F \) represents the Frobenius norm and \( S_R \) the space of matrices \( M \times N \) of rank \( R \). The solution \( \hat{W} \) of (14) is known and obtained by the singular value decomposition of \( W \), after selecting the \( R \) largest singular values.

When the observation matrix \( W \) has a set of missing data, the maximum likelihood estimate of \( \hat{W} \) leads to the generic minimization of (14):

\[
\hat{W} = \arg \min_{\hat{W}} \| (W - \hat{W}) \circ M \|_F, \tag{15}
\]

where \( \circ \) represents the Hadamard product, and the matrix \( M \) of \( M \times N \) is a mask of the know values of \( W \), where \( m_{ij} = 1 \) if the value in known and \( m_{ij} = 0 \) if otherwise.

In the particular case of this algorithm, and given the initial estimate \( \hat{W} \), the EM algorithm estimates in alternate steps:

i. Step E estimates the missing data of the observation matrix \( W \);

ii. Step M estimates the matrix \( \hat{W} \) of rank \( R \) that best approximates the complete data;

The algorithm carries out these two steps until convergence, or until the error stabilizes; in our case until the difference between two successive errors is less than 1x10^-6.

In step E, and considering \( \hat{W}_{k-1} \), the maximum likelihood estimate of the missing data of \( W \) (where \( m_{ij} = 0 \)) are simply the corresponding entries \( \hat{w}_{ij} \) de \( \hat{W}_{k-1} \). We then construct the complete observation matrix \( \hat{W}_k \), whose entries \( \hat{w}_{ij} \) correspond to the entries \( w_{ij} \) of the observation matrix \( W \), if \( w_{ij} \) is an observed value or its estimate \( \hat{w}_{ij} \) if \( w_{ij} \) is missing.

\[
\hat{w}_{ij} = \begin{cases} w_{ij} & \text{if } m_{ij} = 1 \\ \hat{w}_{ij} & \text{if } m_{ij} = 0 \end{cases}, \tag{16}
\]

In matrix notation:

\[
\hat{W}_k = W \odot M + \hat{W}_{k-1} \odot [1 - M]. \tag{17}
\]

On step E, and having the complete matrix \( \hat{W}_k \), the estimates of the missing values, the maximum likelihood estimate \( \hat{W} \) of rank \( R \) is the matrix \( \hat{W}_k \) of rank \( R \) that best approaches \( \hat{W}_k \), in the sense of the Frobenius norm, which is obtained through the singular value decomposition of \( \hat{W}_k \).

Assuming that the optimum characteristic is unknown, this algorithm was applied to each of the possible ranks of the matrix \( Cb \), that is 365.

4.5 B - Unified algorithm of nuclear norm and bilinear factorization

This second algorithm, developed in the paper "Unifying Nuclear Norm and Bilinear Factorization Approaches for Low-rank Matrix Decomposition", has the objective of recovering a matrix \( Z \) of rank \( k \) though a matrix \( X \) with corrupted data, by minimizing:

\[
\min_Z f(X - Z) \tag{18}
\]

subject to rank \( k \).

where \( f() \) is the penalty function. In order to overcome all the problems related to the minimization of the above expression when we have missing data in \( X \), it is rewritten as:

\[
\min_Z f(X - Z) + \lambda \| Z \|_*, \tag{19}
\]

where \( \lambda \) is a trade-parameter between the penalty function and the low rank regularization induced by the nuclear norm.

Without detailing the development of the algorithm, but considering the nuclear norm model to solve the equation above, proposing for this purpose the Augmented Lagrange Multiplier (ALM), it is rewritten as:

\[
\min_Z \| W \odot (X - Z) \|_1 + \frac{\lambda}{2} \| U V^T - \rho^* Y \|_F^2 \tag{20}
\]

that can be done in closed form by the element wise reduction operator \( S_k(x) = \max(0, x - \mu) \), as:

\[
Z = W \odot \left( X - S_k(-X + UV^T + \rho^* Y) \right) + \hat{W} \odot (UV^T - \rho^* Y). \tag{21}
\]

for the cost function L1, or

\[
Z = W \odot \left( X + \rho^* \left( X - UV^T \right) \right) + \hat{W} \odot (UV^T - \rho^* Y). \tag{22}
\]

for the cost function LS.

The implemented algorithm has the following form:

Input: \( X, W \in \mathbb{R}^{M \times N} \), parameter \( \mu \), \( \lambda \), initialization of \( \rho \) 
while not converged do
while not converged do
update \( U = (\rho Z + Y)(X^TV + \lambda L) \) 
update \( Y = (\rho Z + Y)(X^TV + \lambda L) \) 
update \( Z \) (eq. 21 for the penalty function L1 or eq. 22 for LS) 
end while
end while
end while
\( Y = X + \rho (Z - UV^T) \) 
\( \rho = \min((\rho, 10^{-20}) \)
end while
Output: Complete matrix \( Z = UV^T \)
4.6 C - Damped Wiberg algorithm

This method, as described in “Efficient Algorithm for Low-rank Matrix Factorization with Missing Components and Performance Comparison of Latest Algorithms, is in all similar to the previous one, that is to factorize the matrix Y with missing data as the product of two smaller matrices U and V such that:

\[ Y \rightarrow UV^T. \]  \hfill (23)

The Wiberg method eliminates U or V from the problem using the bilinearity of \( f(U, V) \) and applying the Gauss-Newton method to the reduced problem.

Once again, without going in depth on the development of the algorithm, each iteration of the Gauss Newton method is:

\[ u \leftarrow \arg \min f(u, v), \]  \hfill (24)

\[ v \leftarrow v + \partial v, \text{ where } \partial v = \arg \min \|Q_F \partial v - Q_F y\|_2, \]  \hfill (25)

where \( Q_F = I - F (F^T F)^{-1} F^T \).  \hfill (26)

Adding a damping factor to improve the speed of convergence, the final equation used to update \( \partial v \) takes the following form:

\[(G^T Q_F G + NN^T + \lambda I) \partial v = G^T Q_F y, \]  \hfill (27)

\( N \) is a matrix of the form \( N = [N^T_1, N^T_2, \ldots, N^T_n] \) where \( N_j \) is an \( r \times n^2 \) block diagonal matrix having \( r \) sub-blocks \( v_j^T \):

\[ N_j = \begin{bmatrix} v_j^T & \cdots & v_j^T \end{bmatrix}, \quad j = 1, \ldots, n \]  \hfill (28)

The algorithm has the following generic format:

1. Initialize \( \lambda \) with an arbitrary small value (i.e., 0.01).
2. Solve 24 in order to determine \( u \).
3. Verify convergence: stop if it converges, otherwise proceed to 4.
4. Solve 27 in order to \( \lambda \) through Cholesky decomposition and replace on the previous equation.
5. If the value of \( f(u, v + \partial v) \) increases in relation to the previous iteration, define \( \lambda \leftarrow 0.1 \lambda \) and return to step 4, otherwise update \( v \leftarrow v + \partial v \) and \( \lambda \leftarrow 0.1 \lambda \) and return to step 2.

4.7 Results

Shown below are the aggregated results for the three methodologies, considering:

- Observations: considering the entire year of 2013;
- Missing values: the missing positions come from a random number generator with uniform distribution;
- Initialization of missing values: average of the closest values;
- Percentage of missing values: two simulations are performed (A and B) each with 0.1%, 0.5% and 1% missing data in different random positions;
- Ideal rank: rank that leads to the lowest error (Frobenius norm) on the EM method;
- Average error: average of the absolute error
- Relative error: quotient of the difference between the observed value and the estimated value and the latter
- Norm of total error: Frobenius norm of the error of the observed and estimated values

- Norm of missing error: Frobenius norm of the error of the estimated values
- Norm of observed error: Frobenius norm of the error of the observed values

We first emphasise the rank that leads to the minimum error with the EM method; we can confirm through the graphical analysis below that around rank 50 there is a minimum for the error, which also validates the initial assumption that the data is low rank.

![Graphical analysis](image)

**Fig. 11** – Evolution of the error with the rank for EM method, with 0.1%, 0.5% and 1% of missing values

In order to make a critical analysis of the results, these are divided into excellent, with a relative error up to 15% and marked in green, acceptable, with error between 15% and 30% and marked in yellow, and thereafter as unacceptable, marked in orange. The results are summarized in the table below, where the relative error is the average between simulation A and B.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>EM</td>
<td>3.78%</td>
<td>16.45%</td>
<td>8.4%</td>
<td>15.52%</td>
<td>8.5%</td>
<td>4.4%</td>
<td>20.48%</td>
<td>16.31%</td>
<td>4.18%</td>
<td>39.60%</td>
</tr>
<tr>
<td>0.5%</td>
<td>DW</td>
<td>4.12%</td>
<td>16.45%</td>
<td>8.4%</td>
<td>15.52%</td>
<td>8.5%</td>
<td>4.4%</td>
<td>20.48%</td>
<td>16.31%</td>
<td>4.18%</td>
<td>39.60%</td>
</tr>
<tr>
<td>1%</td>
<td>EM</td>
<td>4.12%</td>
<td>16.45%</td>
<td>8.4%</td>
<td>15.52%</td>
<td>8.5%</td>
<td>4.4%</td>
<td>20.48%</td>
<td>16.31%</td>
<td>4.18%</td>
<td>39.60%</td>
</tr>
<tr>
<td>0.1%</td>
<td>EM</td>
<td>3.78%</td>
<td>16.45%</td>
<td>8.4%</td>
<td>15.52%</td>
<td>8.5%</td>
<td>4.4%</td>
<td>20.48%</td>
<td>16.31%</td>
<td>4.18%</td>
<td>39.60%</td>
</tr>
<tr>
<td>0.5%</td>
<td>DW</td>
<td>4.12%</td>
<td>16.45%</td>
<td>8.4%</td>
<td>15.52%</td>
<td>8.5%</td>
<td>4.4%</td>
<td>20.48%</td>
<td>16.31%</td>
<td>4.18%</td>
<td>39.60%</td>
</tr>
<tr>
<td>1%</td>
<td>EM</td>
<td>4.12%</td>
<td>16.45%</td>
<td>8.4%</td>
<td>15.52%</td>
<td>8.5%</td>
<td>4.4%</td>
<td>20.48%</td>
<td>16.31%</td>
<td>4.18%</td>
<td>39.60%</td>
</tr>
</tbody>
</table>

**Table 6** – Relative estimation error for all variables

All methods seem to provide convergence to the observed values. From the 19 variables, 12 have very good estimation results (total energy, energy from negative and positive freezers, energy from interior lights, energy from butcher, interior, exterior supply air and return air temperatures, interior and exterior humidity and air quality), 3 achieve acceptable results (HVAC energy, energy from the bread oven and energy from the negative freezer bread) and 4 unacceptable results (compact machine, energy from exterior lights, energy from the fish shop and percentage of fresh air). Repeating the same analysis but with normalized data, the results in general are worse than the above.

We represent on the tables and charts that follows the complete set of results for some of the most relevant variables, with the intermediate value of 0.5% of missing values. On the charts the observed values are represented in blue, the estimated values in red and the absolute error in green.
**Supply Air Temperature**

Fig. 14 – Missing values estimation results for supply air temperature, with 0,5% missing values, using EM method (left), HUB method (centre) and DW method (right)

<table>
<thead>
<tr>
<th>% Missing</th>
<th>0,1%</th>
<th>0,5%</th>
<th>1,0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>N° missing</td>
<td>11</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>Best rank</td>
<td>78</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Average error</td>
<td>664,02</td>
<td>840,95</td>
<td>682,39</td>
</tr>
<tr>
<td>Relative error</td>
<td>18,70%</td>
<td>21,80%</td>
<td>13,14%</td>
</tr>
<tr>
<td>Error norm.</td>
<td>2,979,80</td>
<td>4,037,90</td>
<td>6,313,80</td>
</tr>
<tr>
<td>Average error</td>
<td>19,340,00</td>
<td>5,987,30</td>
<td>13,583,00</td>
</tr>
<tr>
<td>Relative error</td>
<td>544,79%</td>
<td>155,23%</td>
<td>261,65%</td>
</tr>
<tr>
<td>Error norm.</td>
<td>142,200,00</td>
<td>34,759,00</td>
<td>180,550,00</td>
</tr>
<tr>
<td>Average error</td>
<td>1,030,10</td>
<td>913,72</td>
<td>1,109,30</td>
</tr>
<tr>
<td>Relative error</td>
<td>29,02%</td>
<td>23,69%</td>
<td>21,37%</td>
</tr>
<tr>
<td>Error norm.</td>
<td>4,143,00</td>
<td>4,423,50</td>
<td>10,664,00</td>
</tr>
</tbody>
</table>

**Table 8** – Table with estimation results for HVAC energy

**4.9 Conclusions**

As already mentioned, in general the results are very positive. We emphasise though that the HUB method has worse results that the remaining methods in some specific cases, but the reasons are not investigated. The DW algorithm should give results at least as good as the EM method, which did not occur because we used rank 10 in the simulation parameters, instead of 50, for issues relating to the available memory on the workstation.

The less positive results regarding energy of the compact machine, exterior lights, fish shop and the percentage of fresh air, are detailed next.

The energy of the compact machine has an almost random behaviour throughout the year, so the high error in the estimate was expected; however this is variable of very little relevance, so this result is not highly valued.
The energy fish shop ends up having a relatively periodic behaviour from day to day. Bad estimation results are due to the introduction of new refrigeration equipment in the second quarter of the year and again in the last quarter, which actually introduce a too obvious disruption to achieve good estimation results.

Regarding the percentage of fresh air, due to an equipment problem, during the second half of the year the percentage of fresh air read was constant on 100%, so the estimation results are obviously incoherent and this variable should not have been considered.

The results of the energy consumption from the exterior light is perhaps the most surprising, because it is a variable rather well determined and periodical. The point is that the variable is zero most of the time and, as we determined, seen is weakly correlated with all the other.

With this in mind , and if we look at the estimate considering this variable alone, it appears that there is an average error of 7%, which is considered very good.

Table 11 – Table with estimation results for exterior lights energy, estimated as a stand-alone variable

To conclude, for the global results, the relative error rises slightly with the number of missing values, as expected.

V. CLUSTERING

In this chapter, and outside the scope of this work, we apply a method of clustering in order to determine whether there is interdependence between the different variables, and in particular if it is possible from a small number of days to accurately determine the observations of a specific day. The algorithm used was based on the published work “Sparse Subspace Clustering” from Ehsan Elhamifar and René Vidal.

5.1 Implementation

The method in use is based on the sparse representation to create multiple linear or alike space data clusters, embedded in a larger space. Without detailing the construction of this methodology, in short we have a minimization problem:

$$\min ||c||_1, \text{subject to } y_i = Yc_i \quad (29)$$

where $y_i$ are the observations $(DxN)$ and $Y$ is the matrix obtained from removing the column $i, y_i$.

The optimal solution $c_i$ is a vector while non null values correspond to the column in $Y_i$ that is in the same subspace as $y_i$. The algorithm has the following implementation:

input: a set of points \( \{y_i\}_{i=1}^N \) lying in a union of n linear subspaces \( \{S_i\}_{i=1}^k \)
1. Solve the sparse optimization program (30)
2. Normalize the columns of $C$ as $c_i \leftarrow \frac{c_i}{||c_i||_\infty}$
3. For a similarity graph with $N$ nodes representing the data points. Set the weights on the edges between the nodes by $W = |C| + |C^T|
4. Apply the spectral clustering (min $||a||_1$, subj. to $x=Ba$) to the similarity graph
output: Segmentation of the data: $Y_1, Y_2, ..., Y_n$
After solving (29) for each value of $i$, we obtain a coefficient matrix $C$, with dimension $N \times N$. This matrix $C$ is used to define the graph $G=(V,E)$, where the nodes are the $N$ data and where an edge exists whenever a giver $y_i$ is one of the vectors in the sparse representation of $y_j$ (i.e. $C_{ij} \neq 0$).

### 5.2 Results

Below we have an illustration of the coefficient matrix $C$ obtained from the calculation. It’s very interesting to observe that without any knowledge of the nature of the observations, the algorithm was able to detect 4 “clusters” that correspond exactly to the four seasons in this order; winter, spring, summer and autumn.

![Fig. 21 – Similarity chart for the full range of data (all variables for one year)](image)

Shown below is also a set of graphs representing a few days, selected with criteria, so that is possible to illustrate the relationship between variables. Corresponding days are shown in graphs and the thickness of the connecting line between days indicates if the bond is strong or weak.

![Fig. 22 – Representative grafts of a characteristic winter day in 2 consecutive weeks: 02-01-2013 (left), 09-01-2013 (right)](image)

![Fig. 23 – Representative grafts of a characteristic winter day in 2 consecutive weeks: 03-01-2013 (left), 10-01-2013 (right)](image)

The graphs of the remaining days are very similar, allowing us to conclude that each day may be expressed as a function of another 10 to 15 days. More interesting is taking into account the weight of this relationship, which allows us to conclude that each day can be expressed in terms of only 3 or 4 days that represent 80% of the total weight of the relationship.

### VI. Conclusion and Future Work

In this paper we presented several methods for solving two specific problems; the forecast of the air conditioning energy and the estimation of missing values in the time series of energy and air conditioning data.

To forecast the energy of air conditioning for a certain day, we used the low dimensionality of the data and a linear method to build our forecast. The results obtained demonstrate its applicability, and its simplicity allows its integration in GTC systems.

For the recovery of missing data, three methods were studied, and validated. Here too the low dimensionality of the data and proves that the missing values of almost all variables can be estimated.

Finally, a clustering method was applied, having proved to be possible to represent any one day by a linear combination of other different days.

In short, the proposed objectives at the beginning of this work were achieved.

As future work, we propose implementing HVAC energy forecast method in GTC in order to prove how far it is possible to reduce the energy costs of HVAC, minimizing its consumption in peak hours.

As for the estimation of missing data, we propose to apply this methodology to larger data sets originated from different premises or from a larger timeframe.

The clustering method revealed very interesting results, clearly leaving an open door to determine if there is a set of specific days that allows represent all the remaining.

### References


[9] Anual reports from ERSE