Università
Ca'Foscari
Venezia

Corso di Laurea magistrale (ordinamento ex D.M. 270/2004)
in Informatica - Computer Science

Tesi di Laurea

Optimal statistical designs for energy saving building remodeling

Relatore
Ch. Prof. Irene Poli

Laureando
Vito Capacchione
Matricola 812773

Anno Accademico
2012 / 2013

To my family for their encouragement and patience
To Veronica for her constant support and lost weekends

Acknowledgements

Firstly and foremost, I wuold like to show my sincere gratitude to my supervisor professor Poli Irene who proposed me to work on this compelling and motivating project whereby this final elaborate has been possible. Without her guidance and corrections, certainly I wuold not be so proud of this thesis.
Besides my relator, a grateful thanks to Dr. Borrotti Matteo, Dr. De March Davide and Dr. Slanzi Debora for their concrete support and their crucial suggestions.
This thesis wuold not have been possible unless the support of ECLT research centre (European Centre for living technology) and its staff that provided me experimental data and documentations. A special thanks to Agnese, Cristina, Simona, Stefania, Valentina, Alberto and Alessandro for their support.

I owe my deepest gratitude to my family for their patience. Thanks mum and dad for being full part of my life. A special thanks to my sister, Alessia, for her invaluable presence.

I cannot forget all my friends: Alberto, Enrico, Giacomo, Giulio, Fabio, Marta, Irene, Annachiara, Giovanna, Federico B., Chiara, Mario, Alessandro Ci., Jader, Federico M., Jennifer, Mauro, Caterina, Maicol, Andrea S., Sara, Omar, Davide, Andrea R., Giorgia, Francesca, Alessandro Cr., Christian, Massimo, Davide, Elena F., Elena P., thank you for your authentic friendship.

The following have made available their support in a number of different ways during my academic carrear, I need to list all my fellows: Daniele, Michele, Marco M., Davide, Matteo F., Salvatore, Miriam, Andrea, Marco S., Luca, Enrico, Fabio, Marco F., Filippo, Giulio, Matteo B..

Above all, I wuold like to thank my beloved Veronica who has listened my complaints all these days, tolerating me even when I was beyond all bearings. Thanks for being always next to me.

Contents

List of Figures

Introduction

HVAC system is a set of components whose aim is to provide a comfort indoor environment, in particular it serves heating, ventilating and air conditioning for human comfort in buildings. These systems are mostly installed on large scale buildings such as supermarkets, hospitals and schools. However, HVAC plants are the main responsible for the total energy consumption in buildings. This scientific thesis addresses the problem of reducing building energy consumption preserving optimal comfort levels. The mentioned objectives depends on a set of uncontrollable variables such as room temperature, humidity luminosity and a set of controllable variables such as dimmers, blinds and fan coil. The author's ambition is to find the best configuration of controllable variables in order to maximizing energy savings while maintaining acceptable comfort levels. This research finds huge relevance in environmental issues, which pushed world's government to question about environmental conseguences due to the excessive use of non-renewable energies. In particular energy expenditure due to buildings seems to account for the 30% of the total world energy consumption, as demonstrated in the next chapter. As a result HVAC performance can be definitely improved by setting the best configurations to minimize the energy consumption.
The author's proposal is mainly meant to contribute the scientific research on energy through a statistical approach and to find a practicable solution to a real case study. The case study is based on the 3 offices which belongs to a building situated in a city of the North-East of Italy. The purpose is to find the best configuration of dimmers, blinds and fan coil so as to reduce energy consumption, preserving tolerable comfort levels, in particular to control occupant's comfort have been selected two indexes: PMV and DGI capable of measuring respectively thermal comfort and daylight glare. The control strategy to solve the case study is based on:

- prediction modeling
- optimization algorithm

The prediction modeling is performed through Neural Networks, ARIMAX models and Random Forest so as to predict the HVAC behaviour. To what concern the optimization phase an Exhaustive Pareto Optimization algorithm has been implemented to obtain the best configuration of the controllable variables.
The thesis is organized in four chapters: in the first chapter it is presented an analysis and a description of the environmental problem, the second chapter provides an excursion into the

methodology to solve the problem (state of the art). In particular I focus my attention on the outstanding research performed by Andrew Kusiak who has approached HVAC optimization through data-mining techniques from 2009. The third chapter presents my statistical approach to the case study, it is composed by an explorative data analysis, a feature selection process, modeling and optimization phases. Last but not least the fourth chapter reports conclusions with possible future developments.

Chapter 1

Definition of the problem

1.1 Environmental problem

In the last 30 years energy consumption has grown rapidly, International Energy Agency (IEA) has confirmed it even showing strong relationships among energy consumption, economic developement and population growth. Several and strongly studies [1] state that urban cities are the primary sources of energy consumption. In particular buildings (domestic and commercial) seem to be one of the main consumers accompanied by transports and industries as shown in the following pie chart taken from an IEA research [1] (see technical report [2]):

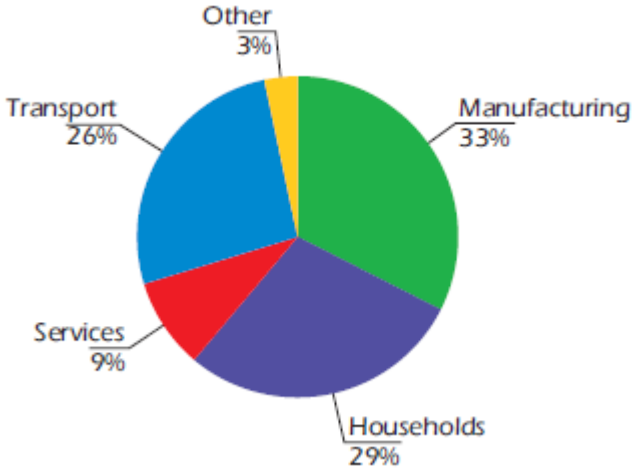


Figure 1.1: Global world energy consumption by sectors in 2005, source from [2]

Taken from an UK work [3] it is possible to see in figure 1.2 the energy consumption with respect to different building types, it is evident as heating ventilating and air conditioning system (HVAC) can be regarded as the main responsible for energy consumption, followed by domestic hot water (DHW) and Lighting. Hvac systems have the purpose to guarantee indoor thermal comfort quality in buildings providing either heating or cooling or just ventilating.
As nowaydays everyone knows the excessive use of energy is causing a heavy impact on envi-

[1] G8 leaders requested to IEA to analyse worldwide trends in energy use and efficency in 2008 in order to sustain the Gleneagles plan of action.

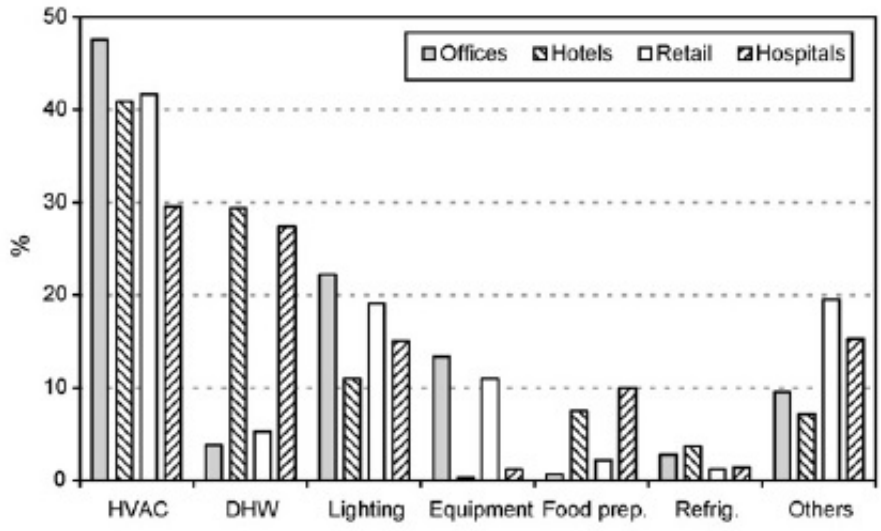


Figure 1.2: Consumption by end use for different type of buildings in UK, source from [3]

ronment, let's briefly recall its effects. Currently fossil fuels (Carbon, oil, natural gas) are the primary exploited resource of energy, they cause several problems to our planet, for instance increment of gas emissions on our air (in particular CO_2), consequently raising the greenhouse effect and acid rains (the latter has an harmful effect on plants, soil, acquatic animals and even on infrastructures). Not to mention deforestation which is supplying CO_2 presence. Drawn from [2], figure 1.3 supports our theory that households highly contributes (1/5 out of total CO2 emissions) to threaten our natural oxygen demand. According to [4], using non-renewable

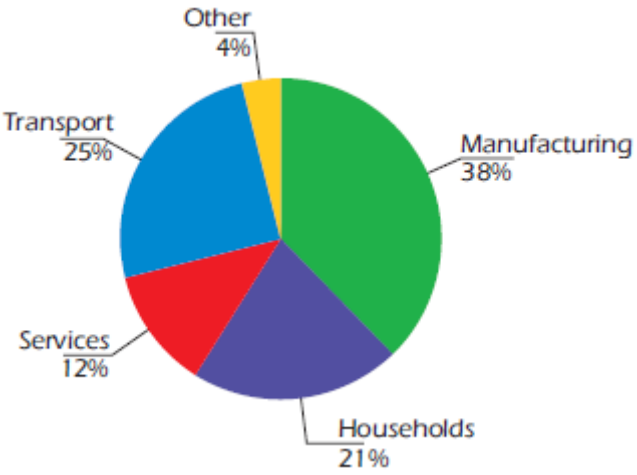


Figure 1.3: CO2 emissions by sectors in 2005, source from [2]

resources at the same current rate they are going to terminate within 2042 apart from carbon reserves which will be available up to 2112. Last but not least air pollution contributes to deplete ozone level in the atmosphere allowing ultraviolet light to get easily our earth.
Being HVAC system the main supplier on building energy consume (confirmed by [3]) I decided

to address my research especially on it.

1.2 World governments commitments

Environmental issues have increased awareness of the problem in governments and public institutions, which emanated national and international directives to struggle with that. The first important step towards an eco-friendly environment was negotiated at Kyoto (Japan) in 1997 by 37 industrilized countries, which committed to reduce greenhouse gas emissions (GHG) by 5% within 2012 against the 1990 levels, this agreement is called Kyoto Protocol. In 2012 part of those countries made a further commitment, proposing to decrease the level by 18% within 2020 with respect to 1990.
European Union's energy policies also moved towards similar objectives, principally on energy savings and use of renewable resources, notably was the 20/20/20 target whereby European Union committed to reach the following goals within 2020 with respect to 1990 levels:

- 20% reduction greenhouse emissions.
- 20% increment of energy consumption from renewable resources
- 20% improvement on energy efficiency

Different instruments have been adopted by EU countries from traditional measures such as fiscal incentives to Tradable white certificates in countries as Italy [2], France, and the UK.
The italian's action plan for energy efficency (PAEE), presented in July 2007 to the European Commision proposed a series of strategies to incentive the energy efficency market in order to reach the 20/20/20 target imposed by European Union. To get those ambitious objectives italian government had not only to revise PAEE (with a 2011 version) but also to approve the National Energetic Strategy (SEN). The four key points of this strategy are:

- Thermal account: promote small-scale intervents to achieve thermal efficency such as building envelope, wall/roof insulations and solar screens and to generate thermal energy from renewable resources. In the same way thermal account contributes to replace old air conditioning systems, heat pumps, boilers with new systems to guarantee higher efficency. These incentives can be obtained by public administrations and private individuals from 2 to 5 years to what extent the project scale.
- tax reduction: a less taxation to supply who is going to do intervents to improve the energy efficency in the extent that type of technology used such as solar panel and biomass heating system.
- Tradable white certificates: energy and natural gas distribution companies every year have to improve energy efficency (hence save energy) to be awarded with white certificates. One certificate corresponds to one tonne of oil equivalent (TOE [3]), it means that if one company

[2] Italy was the first country in the world to apply white certificates, they were approved in July 2004
[3] The TOE unit is commonly used to express large amount of energy 1 TOE = 42 GJ, it derives from the fact that burning 1 tonne of oil releases about 42 GJ

saves 2 TOE, it obtains 2 white certificates. To understand the term tradable I have to say that who does not reach the objectives (let's say saving 5 TOE per year) has to buy certificates from other companies. To summarize either you can produce energy saving by yourself or buy it from someone else that produces for it and you.

- Rotative Kyoto account: it facilitates private and public subjects to obtain loans to invest in systems that allow to reduce gas emissions accordingly to the Kyoto protocol. Private subjects can get 70% loan of the entire investment while public administrations 20% more.

The United States Environmental Protection Agency (EPA) is working on the same wavelength, it constituted several programs and initiatives such as partenership with private sectors to encourage them to reduce their gas emissions, especially EPA ensures to diminish by 25% GHG levels within 2020.

1.3 HVAC system

Nowadays, working in a health and comfort indoor environment like the everyday office is not only fundamental for our productivity and satisfaction but also for that of our company. Let's suppose the HVAC of a supermarket has not been designed accurately how many complaint's letters wuold be delivered to the chief executive officer weekly? Those reasons give us an evident insight of the relationship between our behaviour and the physical environment in which we stand, as a result we can understand the importance of planning no ifs ands or buts HVAC systems.
HVAC system provides heating, ventilating and air conditioning (as expressed by its acronym), it also has to mantain air contamination between tolerable limits, for instance high level of CO_2 in an office has to be managed by the system ([10] is an interesting work that summarizes the acceptable limits of CO_2 and H_2S in indoor environments in the main world countries). Additionally it can dehumidify the environment. Hence, these systems have to include a set of components to guarantee thermal control in buildings. To simplify it as much as possible let's list the basic elements which HVAC is composed by:

- fans: to push in and out air in a building
- supply/return air duct: to drive air in different floors and zones
- supply outlets and return air inlets: to allow respectively incoming and outgoing air in the duct
- filters: to get rid of dust and dirt from air
- boiler: to produce hot water which is passed to the heating coils in order to heat air
- chiller: to refrigerate water which is passed to the cooling coils as to cool air
- condenser: contained in the chiller or sometimes in the cooling tower allow to reject heat as part of the refrigeration cycle

- pumps: to distribute hot/cool water in the system
- heating/cooling coils: to achieve heat or cool air
- damper: to regulate the air flow in and out

The figure 1.4 depicts the mechanism of cooling or heating of an indoor space as an office: On

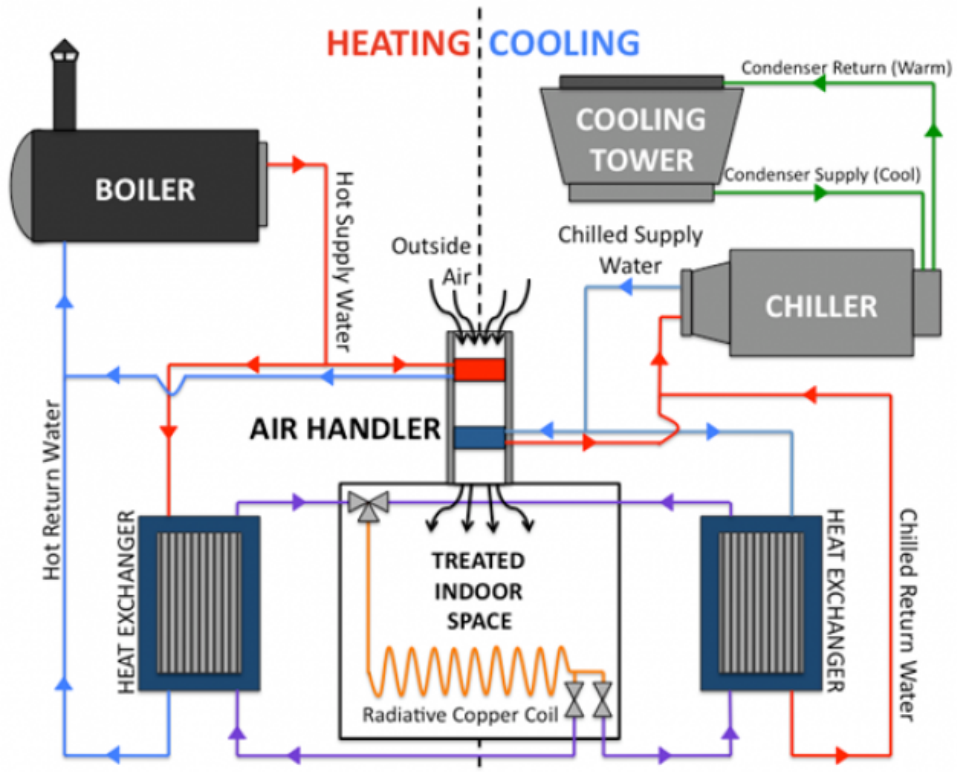


Figure 1.4: Large-scale HVAC system

the right hand-side of our figure there's the cooling area in which the chiller produces cold water through the help of a condeser (contained in the cooling tower) that using a vapor-compression technique allow to waste heat in the outside air and to distribute in the system cold water. The water then flows to the AHU (air handling unit), specifically in the cooling coil (visible in the figure 1.5 on the following page) which is accrosed by air. The latter is cooled and diffuse in the indoor environment. On the left hand-side the boiler has to provide hot water then passed to the AHU (this time it is passed to the heating coil) in order to heat the air that pass through it. In literature the term AHU is a sub-system of HVAC that includes some of those elements previously listed such as dampers, filters,fans, heating and cooling coils (see figure below).

Once the outside supplying air (OSA) gets the AHU device, it enters in the duct through the damper and is immediately filtered to remove dust from air, after that the incoming air is routed towards either heating or cooling coil which are supplied respectively by hot or cool water that allow thermal exchange with the incoming air throughout the thermal convection phenomena. Finally the air is forced towards the treated indoor space (Hot/Cold Deck).
To what concern the exausted air (RA), it is drawn from the indoor space and routed towards outside (EA).

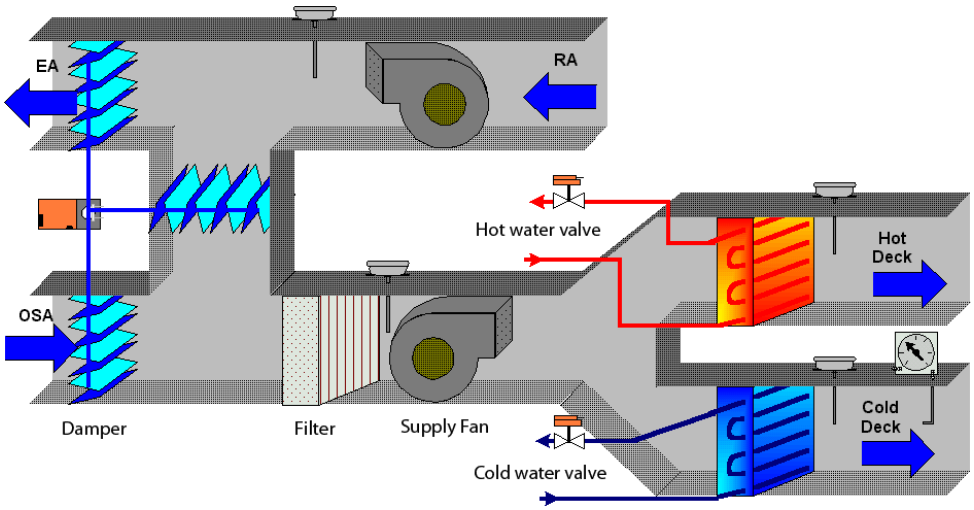


Figure 1.5: Dual-duct AHU system

1.4 Optimization

In this section we address the problem of optimization from a mathematical point of view. Optimization is a field of the applied mathematics, which studies theory and methods to individuate maximum or minimum points of a function. Let's define a function f as:

$$f : X \subseteq R^p \rightarrow Y \subseteq R$$

where the set X is called decision space and the set Y is called criteria space. Given f which depends on $x \in X$ (that can be composed by one or more variables), optimize f means mathematically:

$$\min_{x \in X} f(x)$$

to find the points x that minimize our function f. Optimization means minimization, and also maximization:

$$\max_{x \in X} f(x)$$

This kind of optimization is even known as unconstrained optimization, but it can be also constrained to some equations (constraints):

$$\min_{x \in X} f(x) \qquad subject \quad to \quad g_j(x) \leq h_j \quad with \quad j = 1, ..., n$$

Similarly for maximization:

$$\max_{x \in X} f(x) \qquad subject \quad to \quad g_j(x) \leq h_j \quad with \quad j = 1, ..., n$$

So far we have considered single-objective optimization problems, now I can introduce those called multi-objective problems. The term objective indicate the statement minimizing/maximizing a function. Follow an unconstrained multi-objective problem:

$$\min_{x \in X} f_i(x) \quad with \quad i = 1, ..., m$$

It has the purpose to optimize more functions that could be completely different among them. A very common problem in Economics and Statistics is the constrained multi-objective optimization problem, which is generally described by the following notation:

$$\min_{x \in X} f_i(x) \quad with \quad i = 1, ..., m \quad subject \quad to \quad g_i(x) \leq h_i \quad with \quad i = 1, ..., n$$

The constrained multi-objective optimization problem will be the one that I will use to solve the Str.a.t.e.g.a problem. In the next lines I wuold like to give a wide scenario to what relates optimization. Firstly we can distinguish two kind of problems based on the search space in which we are working on:

- Combinatorial problems

- Continuous problems

To explain them will be sufficient to describe the first category, which contains those problems to determine an optimal solution from a finite set of solutions, in other words we are talking about problems that owns a discrete search space. I recall you the famous "travelling salesman" : given a set of cities and their distances, finds the shortest path so as to reach each city and go back to the beginning, it belongs to the class of combinatorial problems. In continuous problems, variables in the model can get continuous values, commonly real values. To solve continuous optmimization problems, there are some global methods that require the cost function (objective function) to satisfy some properties such as Convexity. While combinatorial problems are often solved through metaheuristics capable of providing nearly optimal solutions. According to literature, metaheuristic methods are abundantly implemented, so I will spend some time talking about them. To mention some of them: Tabu search (TS), Simulated annealing (SA), Particle swarm (PSO), Genetic algorithms (GA) and Ant colony (AC). As identified by their names, their algorithms are inspired by nature, for instance PSO emulates bird flockings and AC reproduces the ants collaboration. Even though these methods were born to work out combinatorial problems, they were extented to face continuous ones. On the contrary to what happen with other algorithms such as descent algorithm or gradients methods, they do not need to calculate derivatives in order to get near the maximum/minimum points. With the purpose to understand the power of the metaheuristics, let's consider to look for the global minimum of an objective function, such that owns the following shape:

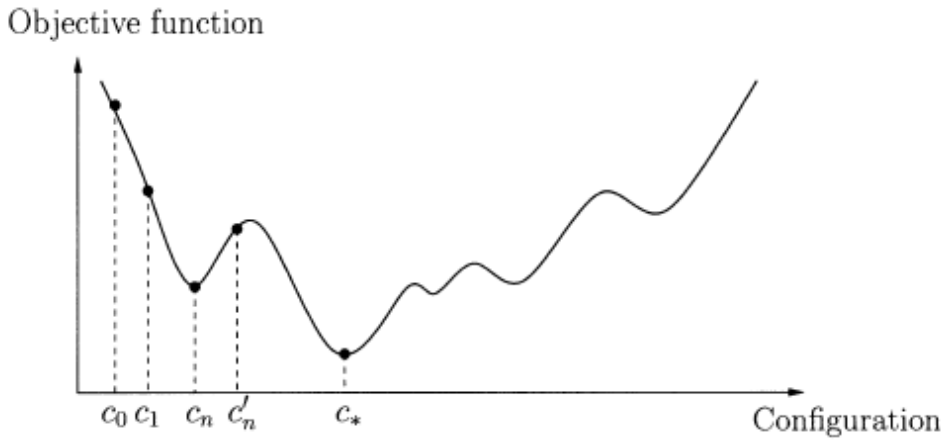


Figure 1.6: Shape of an objective function depending on configurations, source from [11]

A naive iterative algorithm would choose randomly a point, let's say c_1, and other two points one smaller and the other one greater (respectively c_0 and c_n) and wuold select the point that minimizes the objective function between c_0 and c_n. Suppose c_n minimizes the function, it would take the role of c_1 at the next iteration. This algorithm certainly wuold produce a local minimum, but it wuold be the case that it is not a global minimum. This problem results from a wrongly initial selection of c_1, which causes the algorithm to get trapped in a local minimum. Similarly it wuold happen in the most famous first-order gradient descent optmimization algorithm which exploits gradients to get a local minimum. Nevertheless, metaheuristics are structured such that they can get out from a local minimum, this property is based on the idea to degrade the solution some time to time, graphically can be thought as to climb the mountain next to the local point. In order to have a wide scenario in optimization methods, I think can be very usel the taxonomy (1.7 on the next page) provided by [11]. The first discernment has been already explained, it separates combinatorial and continuous problems. Starting from continuous problems, they can be divided in linear and non-linear problems whether their objectives functions and constraints are linear or are not with respect to decision variables. The linear ones are solved using simply linear programming [12]. On the other hand, non-linear problems can be solved in two ways depending on the quantity of local minimum presented by the cost function. Local approaches perfom an optimal work on local search, yet they risk to get trapped in local minimum when the number of local minimum points is high. Earlier introduced, the global methods can be discerned in classical methods and metaheuristic algorithms. The latter can be further branched in distributed and neighborhood classes. GA and PSO are considered distributed since they are population-based, this implies that a set of solutions is managed at each iteration, whereas SA belongs to neighborhood class since it is able to deal with only one solution at each iteration. To what concern combinatorial problems we can distinguish approximate method and the exact one, the latter are able to find the global optimum but they cannot be applied when the size of the search space is too high. Heuristics and metaheuristics can be then placed in approximate methods since get near optimal solutions. Recently a new class has been introduced, it is composed by a mix of metaheuristics and local search in oredr to exploit the advantages of metaheuristics in global search and the advantages of local search, those algorithm has been

defined as hybrids.

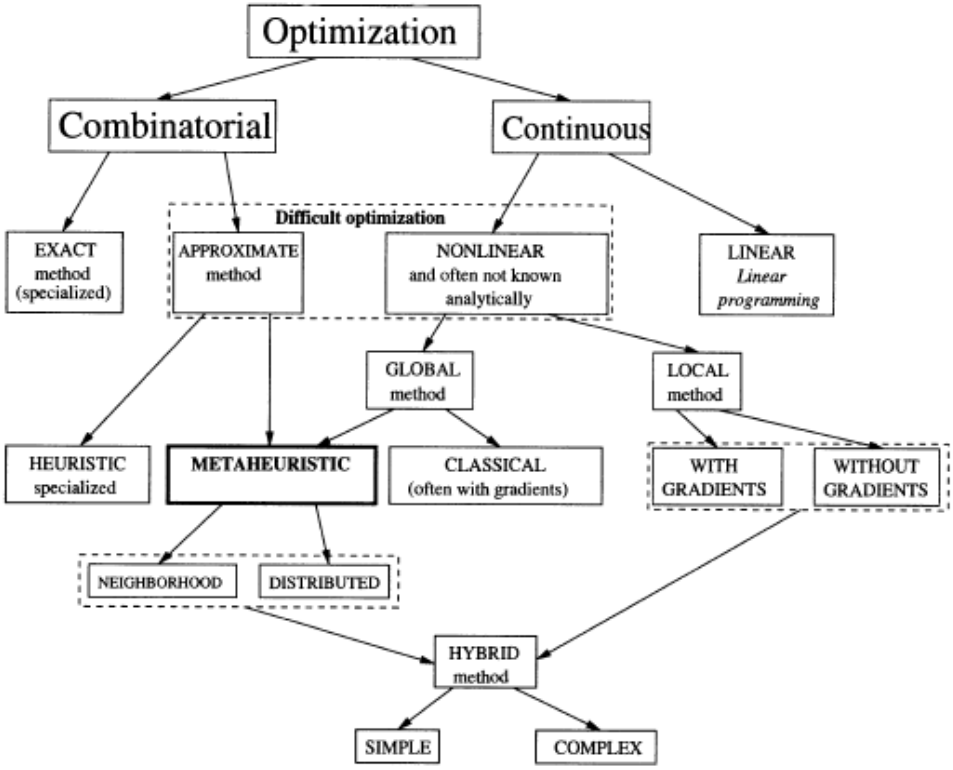


Figure 1.7: Taxonomy of optmization solver, source from [11]

1.5 Stratega project: a case of study

The Str.A.T.E.G.A. project concerns strategies and technological automatizations for efficency and environmental management whose aim is to reduce energy consumption in buildings due to HVAC systems maintaining optimal levels of comfort for inhabitants. This project consider diverse simultaneous objectives such as total electric power and a couple of indexes: predicted mean vote (PMV) and daylight glare index (DGI) to take care of occupants comfort. This reasearch will let me predict the future behaviour of the system and to configure the best settings so as to get the prementioned goal.
The approach used will be a data-driven approach based on a case of study in which data have been collected from 2_{nd} September to 12_{th} December 2013 [4] To what concern the strategy, it will be composed by:

1. statistical prediction modeling
2. optimization algorithm.

In the first step we will simulate the behaviour of the system and test it in potential daily situations. The second step has the purpose of finding the best settings to diminish the energy

[4]Data has been collected by an important company situated in the North-East of Italy, whose main market is based on electronic systems for buildings and environmental technologies.

consumption. As a result the opportunity to predict future behaviour of the system will permit to plan the future control settings conducing to improve energy efficency.

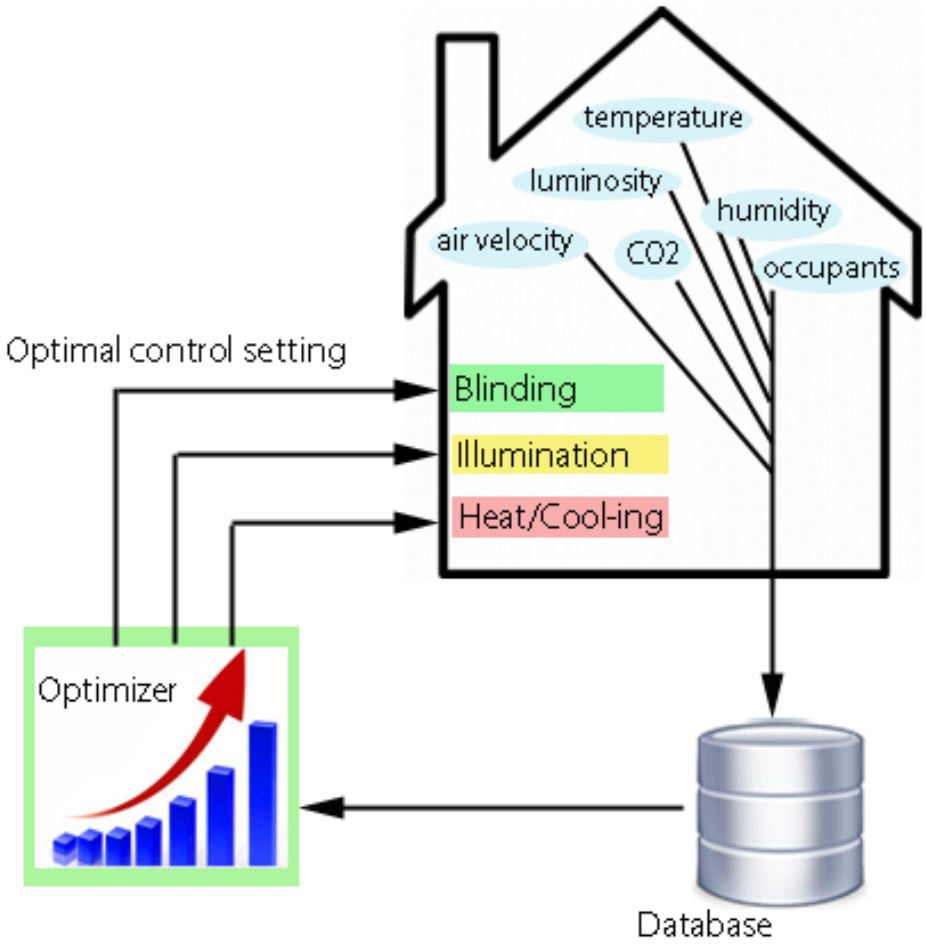


Figure 1.8: Control system

Figure 1.8 shows a scheme of the control system, it gives a general idea of how the optimization energy problem is solved.

1.6 Data description and sensoring system

The building is composed by 3 offices in the same level floor which are called for simplicity OFFICE 1, 2 and 3 (see figure 1.9 on the facing page). In each office has been installed a set of sensors capable of detecting every 5 minutes temperature, luminosity, CO_2 concentrations, humidity, air velocity and presence of people. A further external station have been placed outside the building to get information on external temperature, illuminance, radiation and humidity. After the detection period (about 3 months) the data have been gathered in three different datasets which concern the three different offices. These datasets are formed by 29435 observations and 23 variables.
Each office is controlled by 2 dimmers (identified by d1, d2 during this thesis) for room dimmer-

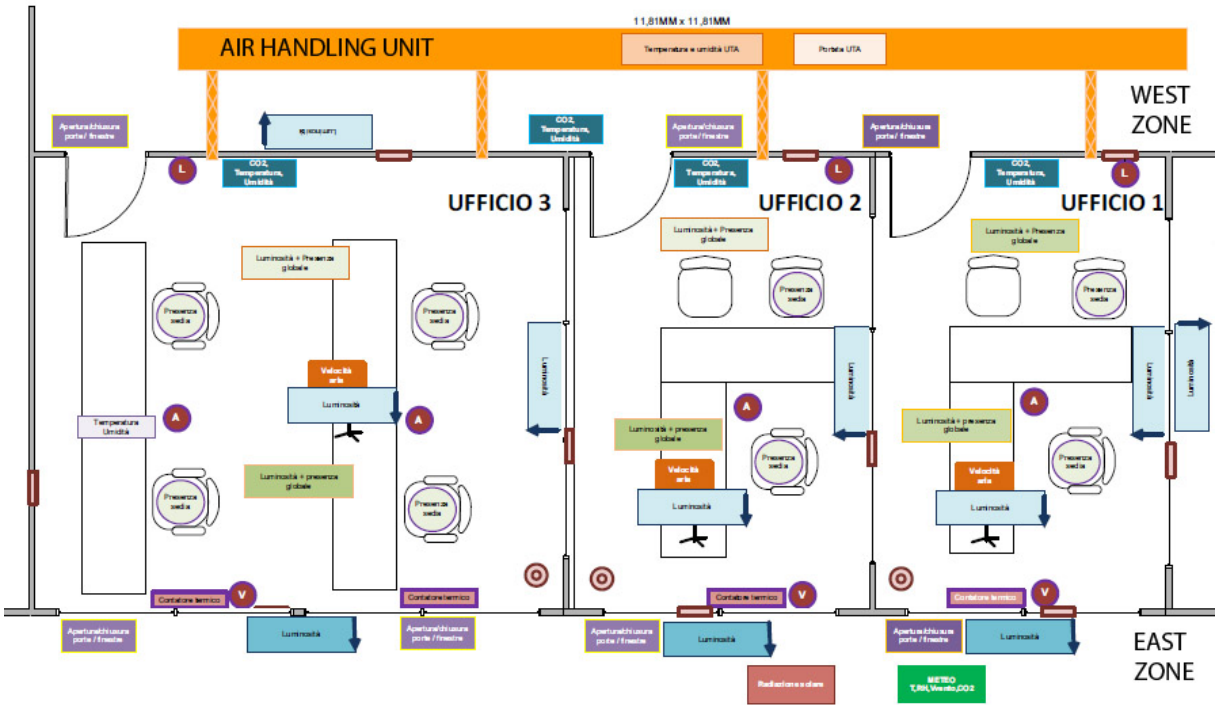


Figure 1.9: Layout offices 1,2,3

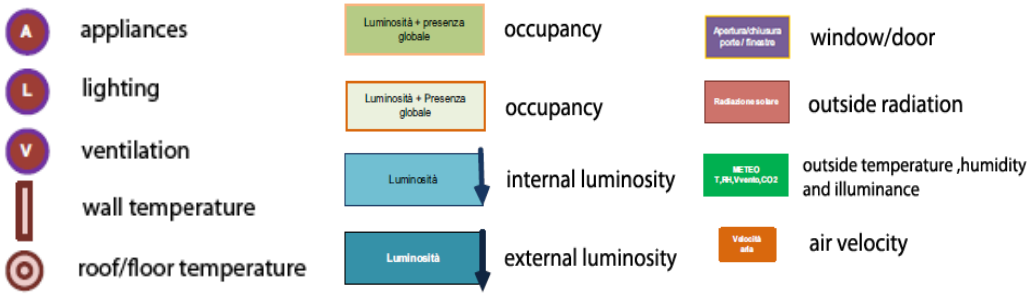


Figure 1.10: offices legend

ization, one blind opening (b) and a fan coil command (f_c). As to facilitate the identification of this group of variables during this paper let's call them controllable variables since they are in fact controllable by inhabitants. Actually they are not completely settable by occupants because the three offices have been thougth to be managed in different ways (automatic or semi-automatic): OFFICE 1 has d1,d2,b manual whereas fc is automatic, OFFICE 2 is completely automatic (d1,d2,b,fc) and OFFICE 3 has only b manual and the other actuators automatic. In table on the next page are shown the values that actuators can assume.

Actuators	id	Levels	Corresponds to
Dimmer 1	d1	0,1	0 Lux and 300 Lux
Dimmer 2	d2	0,1	0 Lux and 300 Lux
Blinds Opening	b	0,50,100	0=open; 100=close
F.C command	fc	0,1	0=off; 1=on

Table 1.1: Actuator values

In each office has been installed internal sensors, let's call them uncontrollable variables which are able to point out temperature (identified by v1), humidity (v2), air velocity (v3), central mean radiant temperature (v4), luminosity west (v5), luminosity east (v6), CO_2 concentrations (v7), occupancy (v8), window sensor (v13), door sensor (v14), corridor temperature (15) and thermal power (y_load). In addition belongs to the same set: outside temperature (v9), illuminance (v10), radiation (v11) and humidity (v12). It is time now to talk about the output variable which is total electric power (y). It is necessary to stress the fact that the latter variable is cumulative, it means that considering the t-th electric power observation, it is calculated summing all the previous observations till t.
The problem, as mentioned in the first section, has to take into account the comfort of inhabitants which is measured by two variables PMV related to the temperature feeling of people and DGI concerning discomfort glare due to the light. Since those two variables are not of common use, I think they need to be examined in depth, however if you are not very much interested in their mathematical calculations, its not knowledge does not preclude the reading of the next chapters. PMV was formulated by Povl Ole Fanger's mind [5], he firstly defined in 1970 the thermal balance equation of human body at [5]:

$$f(M, I_{cl}, v_{air}, t_r, t_a, p_a) = 0 \tag{1.1}$$

Simply to give you a poor explanation, it means that human body is in a thermically wellness condition if there exist a balance between incoming and outgoing energies [6]. As you can see the equation depends on some parameters:

- M : metabolic rate (met[7])
- I_{cl} : clothing insulation index, which measures the thermal clothing resistance (1 clo[8])
- v_{air} : air velocity
- t_r : mean radiant temperature
- t_a : ambient air temperature
- p_a : vapour pressure of water in air

[5]Fanger P.O. was a danish professor at the Technical university of Denmark, he was an expert of thermal comfort and perception of indoor environments.

[6]Incoming energy could be thought as food and oxygen while the outgoing one as the energy that our body needs to exchange with external environment to keep our body around 37 ℃.

[7]1 met corresponds to 58.15 W/m^2 in the International System of measurements (SI)

[8]1 clo corresponds to 0.155m^2 K/W in the SI

Secondly he mathematically expressed PMV index to forecast thermal comfort in indoor environments:

$$PMV = (0.303 \cdot e^{(-0.036 \cdot M)} + 0.028) \cdot L \tag{1.2}$$

where L is thermal load, in other words it corresponds to the exchange energy (heat loss) needed by inhabitants to feel comfortable.

$$\begin{aligned} L \;=\; & \{(M-W) - 3.05 \cdot 10^{-3} \cdot [5733 - 6.9 \cdot (M-W) - p_a] - 0.42 \cdot [(M-W) - 58.15] \\ & -1.7 \cdot 10^{-5} \cdot M \cdot (5867 - p_a) - 0.0014 \cdot M \cdot (34 - t_a) - 3.96 \cdot 10^{-8} \cdot f_{cl} \cdot [(t_{cl}+273)^4 \\ & -(t_r+273)^4] - f_{cl} \cdot h_c \cdot (t_{cl} - t_a)\} \end{aligned}$$

where M is fixed at 1.33 met and W (effective mechanical power) is equal to 0 met, which are considered common values in office as reported at [6].
More common is the relative humidity:

$$UR = \frac{p_a}{p_{sat}} \rightarrow p_a = UR \cdot p_{sat} \tag{1.3}$$

where p_{sat} is the saturation pressure.
To what concern mean radiant temperature, it has been defined as «the uniform temperature of an imaginary enclosure in which radiant heat transfer from the human body is equal to the radiant heat transfer in the actual non-uniform enclosure»[9]. The formulation used to work it out is the one that is based on globe temperature (t_g) and the air temperature (t_a):

$$t_r = [(t_g + 273)^4 + 0.4 \cdot 10^8 \cdot |t_g - t_a|^{1/4} \cdot (t_g - t_a)]^{1/4} - 273 \tag{1.4}$$

Conversely to what happened with t_r which is calculated, t_g has been measured and t_{cl} has been established. Since the first detections have been done in September, the insulation clothing I_{cl} has been initially tuned at 0.5 clo (approximately 0.08 $m^2 \cdot K/W$) which is a tipical value of summer season, whereas during autumn months closer to I_{cl}=1.
Let's finally show the formulation of convection heat transfer coefficient h_c that can be defined as: given two subjects with different temperatures it is the rate in which a subject leaves heat due to their difference of temperature[10]:

$$h_c = \begin{cases} 2.38 \cdot |t_{cl} - t_a|^{0.25} & \text{if } 2.38 \cdot |t_{cl} - t_a|^{0.25} \geq 12.1 \cdot \sqrt{v_{air}} \\ 12.1 \cdot \sqrt{v_{air}} & \text{if } 2.38 \cdot |t_{cl} - t_a|^{0.25} < 12.1 \cdot \sqrt{v_{air}} \end{cases}$$

The next formula expresses an approximation of the clothing surface area factor f_{cl}, which is the clothed body surface divided by the all body surface:

$$f_{cl} = \begin{cases} 1.00 + 1.290 \cdot I_{cl} & \text{if } I_{cl} \leq 0.078 (m^2 \cdot K/W) \\ 1.05 + 0.645 \cdot I_{cl} & \text{if } I_{cl} > 0.078 (m^2 \cdot K/W) \end{cases}$$

To better understand as all these parameters influence our body I really recommend you to read the following scientific work [7]. According to ASHRAE [11] scale, thermal comfort sensation floats between -3 and 3:

[9] Fanger, *Thermal Comfort*, McGraw-Hill Inc., New York, USA, 1970

[10] This process does not need of contact between subjects, what is more the heat loss comes from the warmer element.

[11] American Society of Heating, Refrigerating and Air-Conditioning Engineers

PMV values	-3	-2	-1	0	1	2	3
thermal feeling	Cold	Cool	slightly cool	thermal comfort	slightly warm	warm	hot

Table 1.2: ASHRAE thermal scale

As anticipated by the Fanger's equation, the wellness condition is considered :

$$-0.5 \leq PMV \leq 0.5 \tag{1.5}$$

this is going to be one of our problem's constraints. Concerning DGI, it is a metric which measures discomfort due to glare, it has been developed by Hopkinson[12] in 1972:

$$DGI = 10 \cdot \log \sum_{i=1}^{n} G_i \tag{1.6}$$

where the glare index G_i represents the glare due to each source (our case of study presents only windows) and it is calculated through the formula:

$$G_i = 0.478 \cdot \left(\frac{L_s^{1.6} \cdot \Omega_i^{0.8}}{L_b + (0.07 \cdot \omega^{0.5} \cdot L_w)} \right) \tag{1.7}$$

Let's define one by one the parameters used:

- L_s: is the luminance of each part of the source (cd/m^2)
- L_b: average luminance of the surfaces in the environment, within the field of view (cd/m^2)
- L_w: the weighted average luminance of the window (cd/m^2)
- ω: the solid angle of window (sr stands for steradians)
- Ω: the solid angle of source (sr)

To expatiate more on those variables I put forward you to give a glance at [8]. Being DGI a discomfort metric the higher is its value, more annoying is the glare. Nevertheless light glare can be considered acceptable until value 22, inferior values are irrelevant.

Feeling	DGI Level
intolerable	> 28
uncomfortable	26
comfortable	21
just acceptable	20
just perceptible	< 16

Table 1.3: DGI values

[12]Hopkinson R.O. worked at the Building research station, Garston, Watford, England

To get more information about the latter index I suggest you to see [9]. Let's summarize on tabular 1.4 the uncontrollable and response variables and their unit of measurement.

Uncontrollable Vars	id	unit
Internal Temperature	v_1	℃
Humidity	v_2	%
Air velocity	v_3	m/s
Central Mean Radiant Temp	v_4	℃
Luminosity west	v_5	lux
Luminosity east	v_6	lux
CO_2	v_7	ppm
Occupancy	v_8	binary
Outside Temperature	v_9	℃
Outside Illuminance	v_{10}	lux
Outside Radiation	v_{11}	rad
Outside Humidity	v_{12}	%
Window sensor	v_{13}	binary
Door sensor	v_{14}	binary
Corridor temperature	v_{15}	℃
F.C. th power	y_load	KWh
Comfort Vars		
PMV	x_1	numerical
DGI	x_2	numerical
Response Variable		
Total electric power	y	KWh

Table 1.4: Variables description

Just to give you an idea of what each dataset looks like, it is shown a randomly collection of data related to office 3 in table 1.6 on the next page.

Tempo	d1	d2	b	fc	v1	v2	v3	v4	v5	v6
18/09/2013 00:01	0	0	64	0	24.045	33.73	0	23.3655	16.258	17.897
18/09/2013 00:06	0	0	64	0	24.11	33.76	0	23.3591	16.705	18.195
18/09/2013 00:11	0	0	64	0	24.035	33.76	0	23.3432	16.258	16.258
18/09/2013 00:16	0	0	64	0	24.045	33.81	0	23.3368	17.897	17.003
18/09/2013 00:21	0	0	64	0	24.01	33.84	0	23.3208	17.897	18.195
18/09/2013 00:26	0	0	64	0	24.01	33.97	0	23.308	17.897	18.195
18/09/2013 00:31	0	0	64	0	24.01	33.94	0	23.2953	17.897	17.003
18/09/2013 00:36	0	0	64	0	23.97	34.05	0	23.2857	17.897	18.195
18/09/2013 00:41	0	0	64	0	23.97	34.15	0	23.2698	16.705	17.003
18/09/2013 00:46	0	0	64	0	23.955	34.05	0	23.2602	17.897	16.705
18/09/2013 00:51	0	0	64	0	23.945	34.2	0	23.2506	17.45	17.003
18/09/2013 00:56	0	0	64	0	23.93	34.18	0	23.2314	17.897	18.195

v7	v8	v9	v10	v11	v12	x1	x2	y	y_load
383.6	0	17.9885	0	4.8	61.2679	0.202874464	7.654531098	821.1	2.982525
382	0	17.9885	0	2.6	61.2679	0.211370108	7.181647201	821.1	2.979036667
383.6	0	17.9885	0	1	61.2679	0.199123147	7.087772894	821.1	2.986013333
384	0	17.9885	0	3.2	61.2679	0.20015757	7.560656791	821.1	3.001129444
382.6	0	17.9885	0	4.8	61.2679	0.193674114	7.181647201	821.1	2.980199444
385.2	0	17.9885	0	1	61.2679	0.193205745	7.181647201	821.1	2.970897222
386.6	0	17.9885	0	3.2	61.2679	0.191488556	7.087772894	821.1	2.999966667
386.6	0	17.9885	0	4.2	61.2679	0.185682453	7.181647201	821.1	2.994152778
387.2	0	17.9885	0	3.2	61.2679	0.184615533	7.087772894	821.1	2.976711111
388.2	0	17.9885	0	2.2	61.2679	0.180627749	7.181647201	821.1	2.958106667
386.2	0	17.9885	0	4.8	61.2679	0.179297475	7.560656791	821.2	2.973222778
389.2	0	17.9885	0	3.2	61.2679	0.174819523	7.181647201	821.2	3.001129444

Table 1.5: An hour dataset (12 observations) of the office 3

.

Chapter 2

Methodological approaches to the HVAC optimization problem

Predicting energy consumption in a building is a very complex task, because building energy is affected by simultaneous factors, from engineering aspects such as building structure, walls' thickness, insulation roof to sub-level components like lighting, appliances, fan coils, chiller and boiler. So far, a wide range of proposals have been presented to forecast and optimize thermal comfort and energy expenditure of offices, public buildings and real estates. Some of those proposals related to energy prediction phase have been recaped at [13].
The current chapter is thought to prepare the reader to analyse the next chapter equipped with the appropriate instruments used in literature to solve similar problems. Obviously, will be presented only those works which I personally consider the most significant.
In literature, the developed models can be branched in three different categories:

- engineering methods
- statistical methods
- artificial intelligence methods

The first approach will be just introduced because of the different academic background owns by the author of this thesis, conversely I will get in depth to what concern the second and third methods.
Model building phase is then followed by the optimization step, in this chapter I will present only meta-heuristic algorithms of the following two types:

- evolutionary programming
- swarm intelligence

2.1 Engineering approach

Even though this approach is certainly the most accurate in thermal and energy predictions, it involves the application of physical, mathematical and thermal dinamics principles that yields

their modeling quite though. International Standard Organization (ISO) formulated elaborated standard energy calculations that can be looked up at [14].
Nowadays, in practice, all those formulations are past, and substituted by decades of software tools that have been developed to simulate building consumptions such as DOE-2, BLAST, TRNSYS and EnergyPlus. A complete list can be seen at the website of U.S department of Energy. Despite using those useful tools make the professionist'life easier, they request for meticulous information concerning the building structure, weather conditions and sometimes these are difficult to retrieve.
Usually the first approach is called Divide et Impera, it has the purpose to reduce the complexity of the general energy consumption, decomposing it in easier submodels, therefore it achives the energy demand of sublevels as chiller, boiler, lighting. Finally it gather all the submodels: summing all them up in order to explicit the total electric power. For instance, this approach has been used by Yao and Steemers at [15]. According to Al-Homoud at [16], if the objective is to analyse trends and maybe compare systems, the use of simplified models is acceptable as performed by Wang and Xu at [17], which reduced the complexity of phisical characteristics of the building considered simply using phisical details and information based on frequent values. In addition they determined the most suitable model parameters implementing a GA.

2.2 Statistical approach

In predicting models, a motivational work has been performed by Mustafaraj, Chen and Lowry [18]. They built simple linear parametric models capable of forecasting daily internal temperature and humidity with time scales of 30 minutes and 2 hours. Their data were collected from a Building management system (BMS) installed on an open-plan office at the second floor of the Portman House in London between 2005 and 2006. BMS was set to record internal temperature and humidity, supply air flow-rate, supply air temperature, supply air humidity, outside temperature and humidity, chilled and hot water temperature. The three authors used the following models:

- Autoregressive Integrated Moving Average (ARIMA sometimes inappropriately called Box-Jenkins)
- Autoregressive with external input (ARX)
- Autoregressive moving average with external input (ARMAX)

To evaluate their accuracy in prediction, has been used four very common performance metrics:

- Goodness of fit (G) $G = \left(1 - \frac{\sqrt{\sum(\hat{y_i}-y_i)^2}}{\sqrt{\sum_{i=1}^{N}(y_i-\frac{1}{N}\sum_{i=1}^{N}y_i)^2}}\right)\cdot 100$ for i = 1,...,N
- Mean square error (MSE) $MSE = \frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y_i}|^2$
- Mean absolute error (MAE) $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y_i}|$
- Coefficent of determination (R^2) : $R^2 = 1 - \frac{\sum(y_i-\hat{y_i})^2}{\sum(y_i-\overline{y})^2}$

where N is the total number of observations, y_i are the observed values, $\overline{y}$ is the sample mean [1] and $\hat{y}_i$ are the predicted values from the model. Let's just show the most compelling validation results related to room temperature and humidity in autumn season:

Step ahead	G	R^2	MSE	MAE
6	80-85	0.95-0.98	0.008-0.01	0.072-0.092
12	74-77	0.92-0.95	0.013-0.024	0.093-0.116
24	65-70	0.88-0.91	0.03-0.05	0.12-0.17
Simulation	55-60	0.8-0.87	0.035-0.055	0.195-0.22

Table 2.1: Validation criteria of internal temperature using ARIMA in autumn season

Step ahead	G	R^2	MSE	MAE
6	79-83	0.94-0.97	0.008-0.0013	0.081-0.093
12	68-77	0.92-0.94	0.015-0.03	0.095-0.119
24	58-64	0.83-0.91	0.04-0.06	0.13-0.19
Simulation	50-55	0.75-0.78	0.045-0.063	0.21-0.235

Table 2.2: Validation criteria of internal temperature using ARMAX in autumn season

Step ahead	G	R^2	MSE	MAE
6	68-79	0.92-0.96	0.85-1.5	0.42-0.65
12	65-74	0.9-0.93	1.25-1.8	0.55-0.75
24	57-65	0.8-0.87	1.78-2.25	0.73-0.8
Simulation	51-55	0.77-0.81	2.3-3.15	0.85-1.2

Table 2.3: Validation criteria of internal humidity using ARIMA in autumn season

Step ahead	G	R^2	MSE	MAE
6	73-77	0.93-0.95	0.91-1.2	0.45-0.51
12	61-69	0.89-0.91	1.4-2.45	0.62-0.78
24	54-60	0.78-0.84	2.5-2.85	0.78-0.85
Simulation	48-52	0.74-0.79	2.95-3.5	0.87-1.3

Table 2.4: Validation criteria of internal humidity using ARMAX in autumn season

To be precise those values contained in tables 2.2 and 2.4 are actually confidence intervals related to the performance metrics. I wuold like now to summarize some observations expressed by the authors, which could let me reflect:

- Despite applying phisical and mathematical models could give better results, they involve the knowledge of very complex relationships among all the parameters. As a result simply using linear parametric models, it is possible to obtain acceptable forecastings, see figure 2.1.

[1] $\overline{y} = \frac{1}{N}\sum_{i=1}^{n} y_i$

This is the main reason why I will apply this kind of approach, rather than an engineering one (chapter 3).

- The higher is the time scale prediction (step ahead), the lower will be the accuracy as shown by the previous tables, so in my case of study I have to take into consideration this fact in order to avoid too long scale predictions which could bring me to erroneous predictions.

- ARIMA (identified by BJ in figure 2.1) outperforms the other models, furthermore their data fits better to temperature models rather than humidity ones, then comparing more models is a rule of thumb if I want to get good models for my problem.

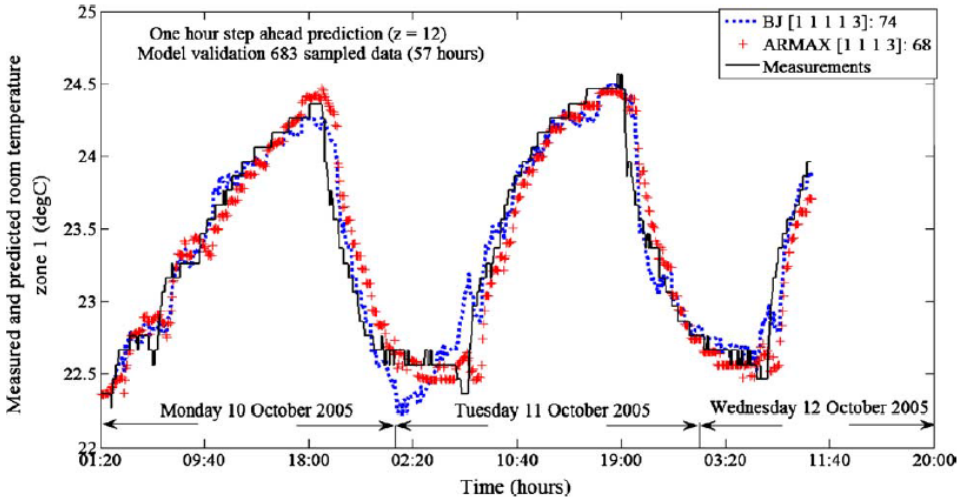


Figure 2.1: Goodnes of ARMAX and ARIMA (BJ) models with respect to real measurements from 10th to 12th October 2005, source from [18]

Practically these models could be integrated to PID controllers of schools, hospitals and public buildings to regulate in advanced the thermal comfort settings guaranteeing always acceptable comfort conditions.
Different models of time series have been implemented in literature for modeling building energy, for instance an ARIMA model was perfomed well at [19] to predict online energy demand of an air conditioning system, while an ARIMAX model was applied at [20] to regulate peak demand of commercial building energy.
Sometimes changing the point of view of the problem could give interesting insights, for instance Ooka and Komamura [21] planned an optimal design method for building energy systems, they used a variant of a genetic algorithm to provide the best combination of equipment capacity and and best operational planning for heating, cooling and power with respect to energy consumption and CO_2 emissions. The gist of their research work can be summarized by the following statement: in designing phase find the best combination and operation of HVAC equipment, means finding the best conditions to maximize energy savings. Data has been simulated from "Computer Aided Simulation for Cogeneration Assessment & Design" application (also said CASCADE). This work

proposes an optimal design method, let's see graphically the flow process (figure 2.2 on the facing page):

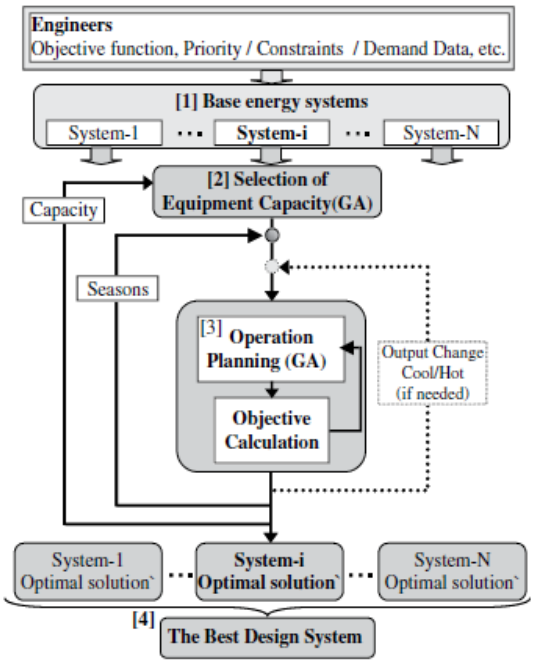


Figure 2.2: Process flow followed by Ooka and Komamura, source from [21]

The flow can be divided in four main parts:

1. Individuation of the potential basic energy systems.

2. Optimization of equipment capacity of each energy system.

3. Optimization of the operational process of each energy system.

4. Choosing the best designed system among the optimal solution candidates.

Firstly, enumerate the possible energy systems (see figure 2.3 on the next page), which are considered the candidates to the final design, the second and the third step are performed concurrently and in this phase the Multi-Island Genetic Algorithm (MIGA) is implemented, while the final step determine the optimal system choice based on the minimization of energy consumption and CO_2 emissions.

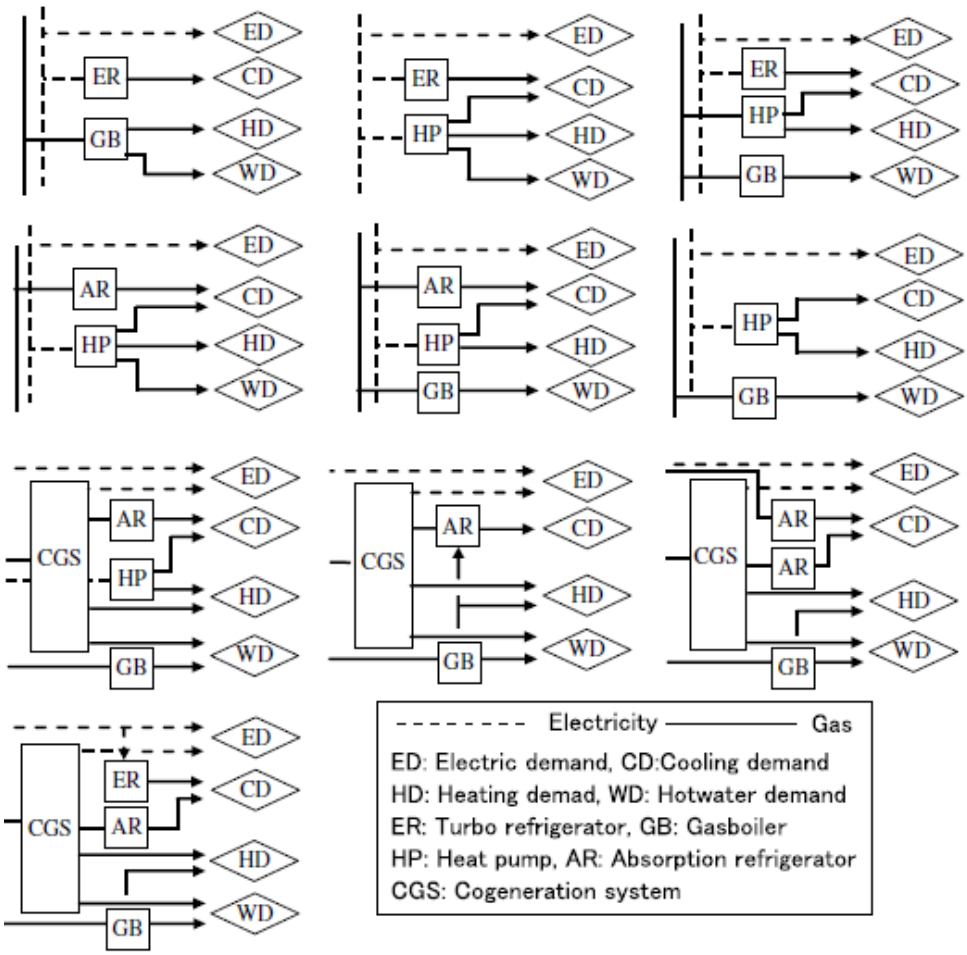


Figure 2.3: Diagrams of some potential basic energy systems, source from [21]

To apply a genetic algorithm each candidate is characterized by a binary sign row called chromosome, in this case each individual is described by two chromosomes, one to define the equipment capacity and the other one to identify the operational process (figure 2.4 on the facing page). The latter chromosome is a time series of output coefficents (operational load factor).
To better understand the binary string let's look the first 3 bits of the first chromosome, which refers to the capacity of the turbo refrigerator (ER), which is labelled as 100 that converted to decimal numbers corresponds to 200kW. Hence, following the same conversion for all the string, the first chromosome spot the candidate design system that owns the following capacities:

$$(\text{ER} = 200\text{kW} \ , \ \text{CGS} = 120\text{kW}, \ \text{HP1} = 75\text{kW}, \ \text{HPw1} = 160\text{kW}, \ \text{HPw2} = 160\text{kW})$$

For operation control, the chromosome represents the 24 hour output coefficients for each equipment, each three-bits fragment can assume values as 000,001,010,011,100 and 101 that corresponds to the output coefficients 0, 0.2, 0.4, 0.6, 0.8 and 1.
Differently from standard GA, MIGA is a distributed algorithm and it is characterized by the fact that at each generation the population is divided in sub-populations called islands, at which are applied the genetic operations: selection, crossover and mutations. A key feature of this GA

is the migration operation which allow candidates in the population to periodically change island (as shown in figure 2.5 on the next page).

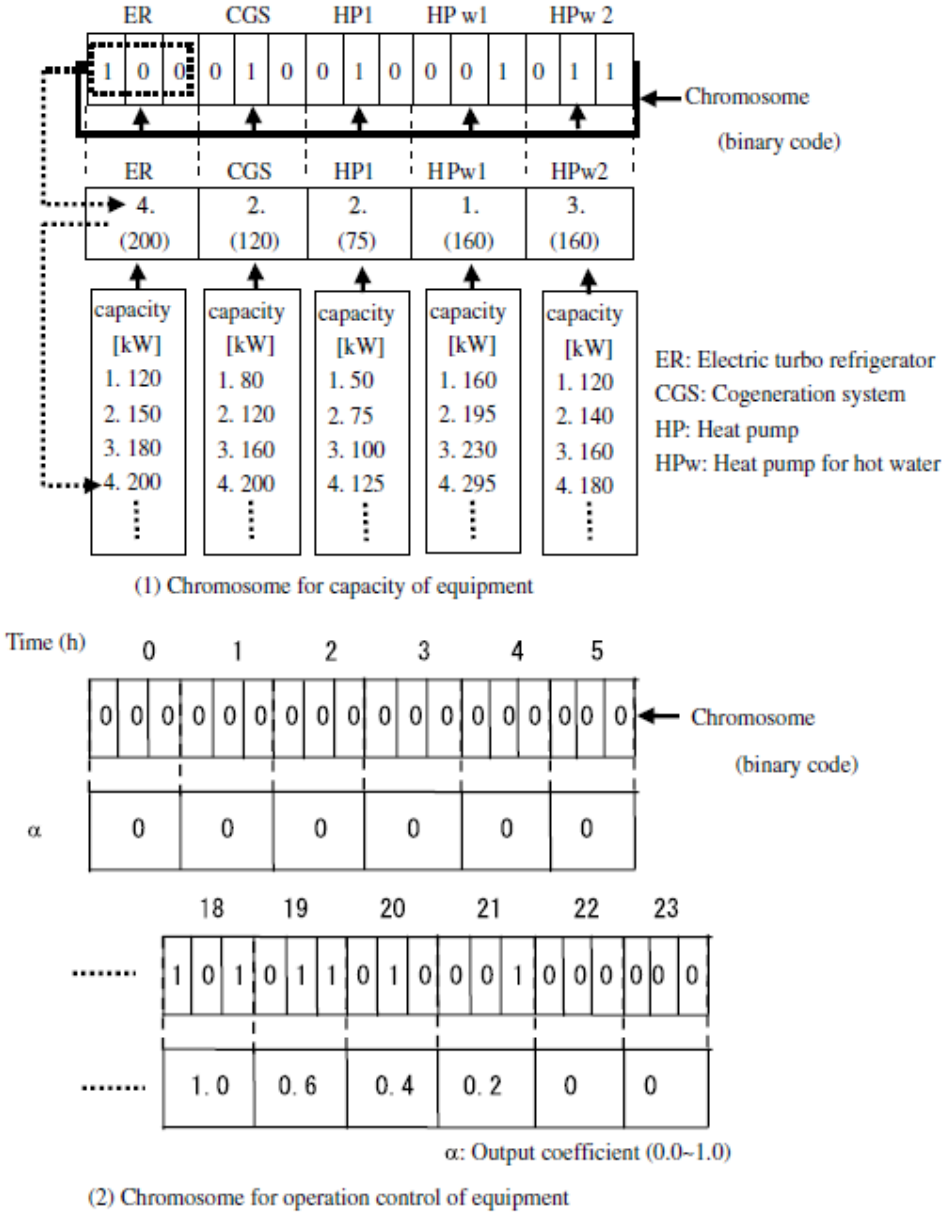


Figure 2.4: Chromosomes for capacity and operation of equipment, source from [21]

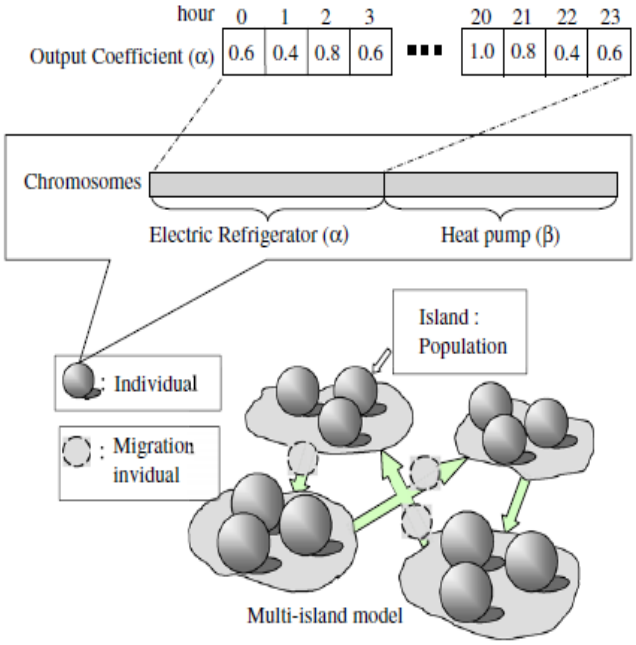


Figure 2.5: Islands of MIGA algorithm, source from [21]

As shown by the authors MIGA worked quite well, however in terms of time computing, MIGA has requested several hours to foresee a single day estimation in a study case, thus it wuold involve a couple of months to estimate a year, and this can be considered a drawback.

2.3 Artificial intelligence approach

A very common machine learning exploited in building energy in the last decades is Artificial neural network (ANN), this is due to its extraordinary facility to model very complex non linear problems. Just to mention some works: Kalogirou at [22] used back propagation algorithm to forecast the demanded heating load of 225 buildings, while at [23] were employed recurrent neural networks to predict hourly energy expenditure as a result of heating and cooling needs. To note the study performed by Magnier and Haghighat [24] in 2010, which cope with an optimization problem very similar to mine one, in particular the two objective functions were:

1. $F_1 = |PMV|_{AVG} \cdot (1 + \frac{N_{dis}}{100})$
2. $F_2 = (E_{heat} + E_{cool} + E_{fan}) \cdot (1 + \frac{N_{dis}}{100})$

to assess respectively thermal comfort and energy consumption. From F_2 it is possible to evince as the authors expected energies of heating E_{heat}, cooling E_{cool} and fan E_{fan} to significantly change the total energy demand, whereas lighting and appliances are not considered relevant. In F_1 there are $|PMV|_{AVG}$ which is the absolute annual PMV average and N_{dis} is the number of hours where $|PMV| > 0.5$ (that means occupants dissatisfaction). This study was conducted to find the best design (HVAC system settings and thermostat programming) so as to optmize the two objective functions. The dataset was generated using a simulation environment, called

TRNSYS, taking special care of the geometry and materials of a residential house[2] placed at the NRC campus. This database as shown by figure 2.6, was used to feed an 1-hidden layer neural network composed by 20 input variables and 5 output neurons ($|PMV|$, N_{dis}, E_{heat}, E_{cool}, E_{fan}). To solve the optimization problem has been used a metaheuristic Genetic algorithmm, that allow

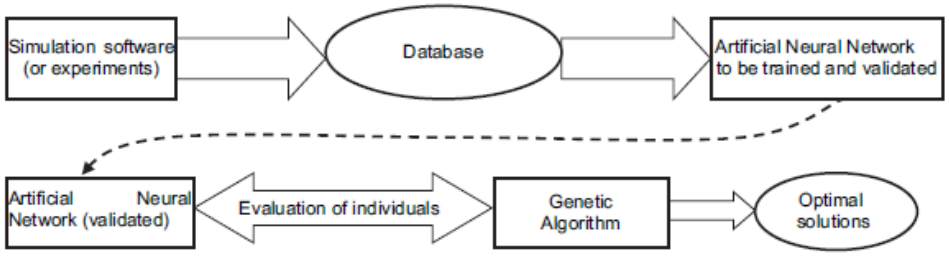


Figure 2.6: Flow process followed by Magnier and Haghighat, source from [24]

the authors to get the optimal front represented in figure 2.6:

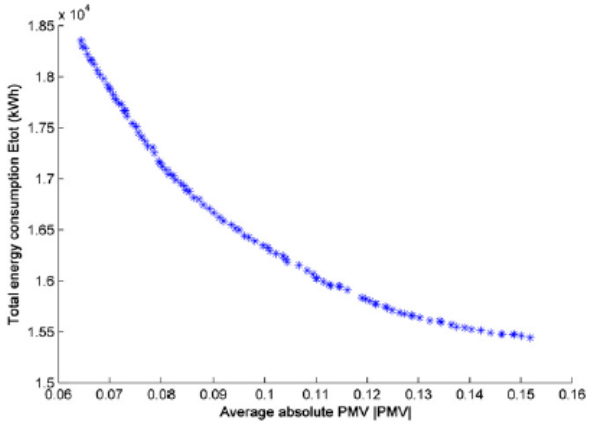


Figure 2.7: Optimal front, source from [24]

The large amount of spreading solutions in the optimal front related to one year predictions, report very different ways to design the system. In terms of time employed, if they had not used the ANN, but they had used the GA directly on TRNSYS model, they wuold have crashed with an unfeasible problem. Last statement is resonable thanks to the application of the Response Surface methodology, (that is actually summarize in figure 2.6), where the ANN permits to predict only the potential evolutions of the system and GA chooses the optimal solutions among those, whereas applying GA to all the possible evolutions (database), the execution time wuold have been unpracticable.
Alternatively, another approach sometimes used and already introduced by the engineering approach, is the one of selecting sub-components of the main problem and optimize those, for istance, Chow, Zhang, Lin and Song [25] studied optimal controls of electricity and fuel in

[2]the mentioned residential house is one of the two Twin Houses placed at the National research council campus in Canada

absorption chiller system, taking special care of economical question. They summarize the objective function as follows

$$\phi = C_f \cdot Q_f + C_e \cdot (P_{chiller} + P_{cool}) \tag{2.1}$$

where ϕ is the total energy cost rate (\$/hour), C_e and C_f are the unit electricity and fuel cost respectively (\$/KWh and \$/Kg), Q_f is the fuel consumption rate (Kg/hour), while $P_{chiller}$ and P_{cool} are the chilled water and cooling pump electric powers (KW). The authors wanted to minimize the cost function ϕ, finding the best combinations of the following vector

$$(Q_c,\ m_2,\ m_3,\ T_2,\ T_3)$$

where

- Q_c: cooling load operation (uncontrollable variable)
- m_2: chilled water mass flow rate
- m_3: cooling water mass flow rate
- T_2: chilled water temperature
- T_3: cooling water temperature

Using a crossvalidation[3] approach on a case study, was possible to choose the best parameters for a Multi-layer perceptrons architecture, composed by 2 hidden layers, 5 neurons for the input layer and 4 neurons for the output one, follow its structure:

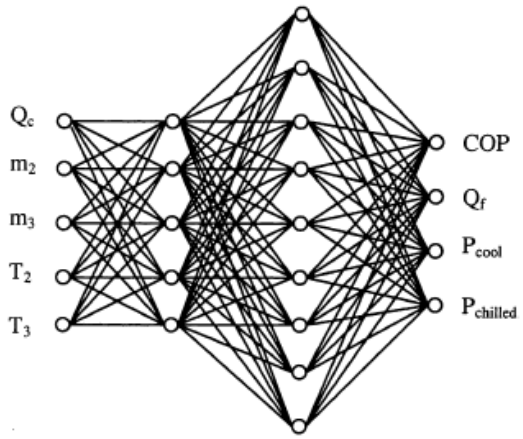


Figure 2.8: Multilayer perceptrons model of the absorption chiller, source from [25]

Even though wuold have been sufficient just three output neurons (see equation 2.1), the authors decided to add the coefficent of performance (COP) to have an indicator of the efficency

[3]Cross-validation is a statistical technique that allow you to evaluate or compare different models, it can be used when you have a large amount of data, it divides the dataset in k approximately equal parts (it is said k-fold crossvalidation), one of these is used to validate each model, while the other k-1 subsets are used to feed the models. It is always accompanied by some performance metrics to evaluate the accuracy of each model. Thanks to these metrics, it is possible to choose the model that fits better the dataset.

of the plant.
After the ANN model, a standard GA was integrated to the process, let's show the flow chart:

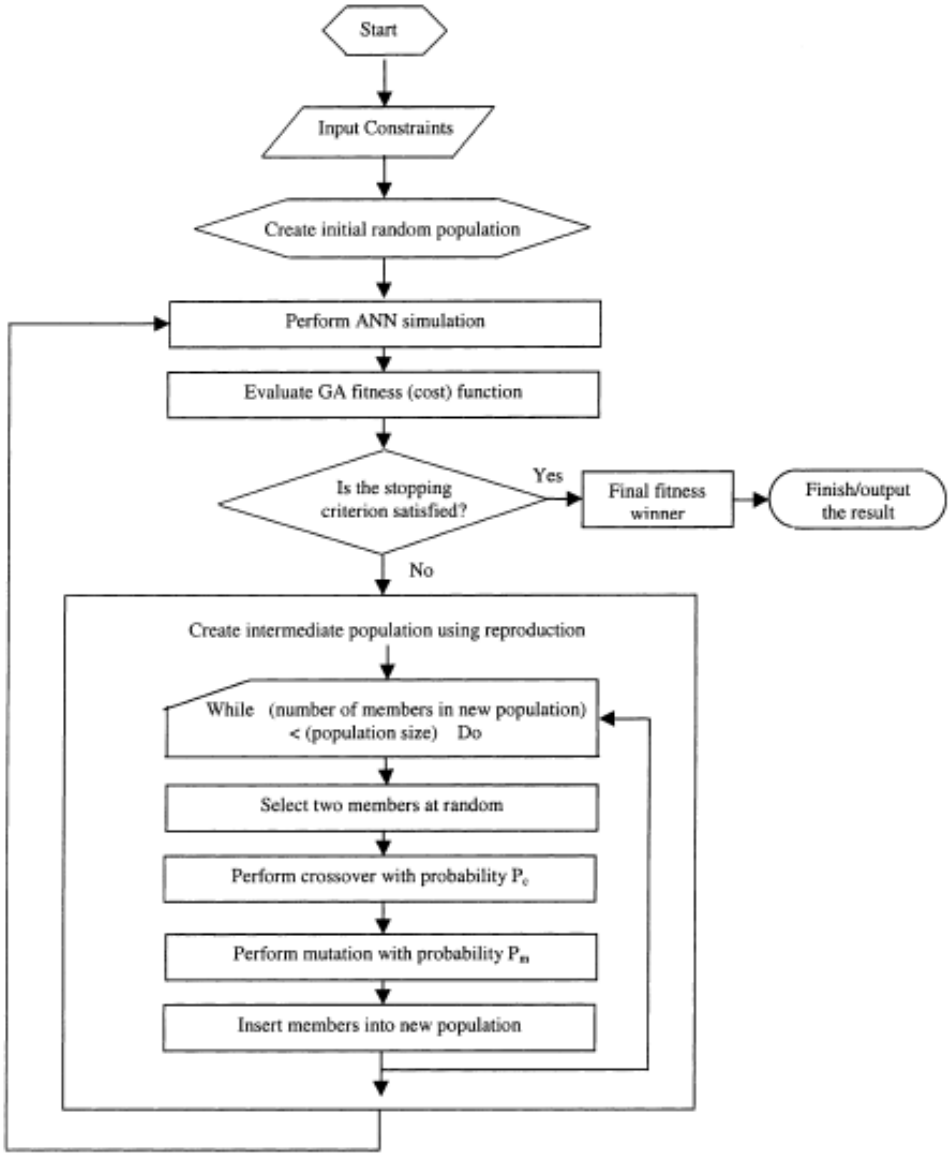


Figure 2.9: Optimization flow process of ANN and GA, source from [25]

The initial population was chosen randomly, and of size 50, each set of control variables (Q_c, m_2, m_3, T_2, T_3) is single out by a chromosome of size 80 bits (to allow the 5 input variables to get every possible value). In order to obtain the next generation, was firstly used a selection operator based on roulette-wheel, secondly a single point crossover with propability to be applied equal to 0.5 and thirdly a mutation operator bit-by-bit with a very low probability to be activated equal to 0.033. These three genetic operators are applied for 100 generations.
Finally the researchers tested their work, alternatively fixing constant the control variables, their results were, as expected, the more control variables you allow in your control strategy, the higher

is the energy savings.
In my opinion the most interesting research was developed by Kusiak [26], [27], [28], [29], [30] who has concentrated his scientific work on modeling and optimizing building energy consumption and thermal comfort since 2010. The reason why I focused my attention particularly on his work during my research period is that he followed a meticolous statistical data-driven design approach and this allows him to get easily outstanding results in energy efficency. Let's try to summarize his reasearch highlighting the key points.
The large amount of data that he collected in these years were conducted at the Energy Resource Station (ERS[4]) by the Iowa Energy Center. His datasets were characterized by high dimensionality (approximately 300 variables) and observations recorded at 1-min sampling intervals for several months and in different seasons. Once obtained the dataset he derived a new one expressed in 30-min intervals and divided it in three subsets with the purpose to exploit them respectively for training, testing and validate his models.
After the initial pre-processing phase, a feature selection was applied in order to select the most significant variables that will be used for the model building stage of energy consumption, room temperature and humidity (his responses variables). To extract the most relevant features from the domain knowledge he performed the k-fold cross validation integrating several wrapper algorithms such as Linear regression, Pace regression, Support Vector Machine (SVM) regression, Multiple Layer Perceptrons (MLP) and Boosting Tree. Each time one of these wrappers selected one variable, the latter obtained a vote. Eventually the most voted were chosen to contribute to the models. The 12 most voted variables represent those which have the greatest impact on the model (see figure 2.10).

	Wrapper = Linear regression + Genetic search	Wrapper = Pace regression + Linear forward	Wrapper = Linear regression + Greedy	Wrapper = SVM regression + Genetic search	Wrapper = MLP + Genetic search	Rank
	No. of times selected	No. of times selected	No. of times selected	No. of times selected	No. of times selected	Total
Internal load ($t-1$)	10	10	10	10	10	50
SA set point	10	10	10	9	10	49
MA-Temp	10	10	10	10	9	49
OA-Hum	10	10	5	10	8	43
SPSP set point	10	10	10	3	9	42
BAR-Pres	10	10	7	10	4	41
OA-Temp	3	10	8	8	10	39
SOL-Beam	10	7	5	10	7	39
Internal Load (t)	5	5	4	9	10	33
SA-Hum	9	9	4	6	3	31
IR-Rad	8	9	5	2	5	29
SOL-Hor	2	7	5	3	10	27
WIND-Dir	8	5	5	7	1	26
WIND-Vel	2	8	9	1	5	25
OA-CO_2	0	2	2	3	4	11

Figure 2.10: The 12 most significant ranked variables obtained by a 10-fold cross validation technique, source from [28]

The second main step was the model construction based on the following regression techniques: Multiple Layer Perceptrons ensemble (MLP), Chi-squared automatic interaction detector (CHAID), Boosting Tree, Random Forest, SVM, Regresion tree (C&RT) and Multivariate Adaptive regression spline (MARSpline). To train these models were used the training subset, while to

[4]ERS is a facility research for testing of commercial HVAC systems, located at Iowa City, Iowa, US.

test them the testing set. As to validate his models, Kusiak used four very common performance metrics:

- Mean Absolute Error
- Standard deviation Mean Absolute Error
- Mean Absolute Percentage Error
- Standard deviation Mean Absolute Percentage Error

These metrics can be calculated as follows

$$MAE = \frac{\sum_{i=1}^{n} AE_i}{N} \quad with \quad AE = |\hat{y} - y| \tag{2.2}$$

$$APE = |\frac{\hat{y} - y}{y}| \tag{2.3}$$

$$MAPE = \frac{\sum_{i=1}^{n} APE_i}{N} \tag{2.4}$$

$$Std_{AE} = \sqrt{\frac{\sum_{i=1}^{n}(AE_i - MAE)^2}{N-1}} \tag{2.5}$$

$$Std_{APE} = \sqrt{\frac{\sum_{i=1}^{n}(APE_i - MAPE)^2}{N-1}} \tag{2.6}$$

where AE in 2.2 is the absolute error, N is the number of observations used for validate, $\hat{y}$ is the predicted value and y is the measured value. As shown by figure 2.11 it is possible to evidence as MLP ensemble outperforms the other models, this is why MLP ensemble was chosen to model energy consumption y_1, room temperature y_2 and humidity y_3.

Dataset	Algorithm	MAE	Std_AE	MAPE (%)	Std_MAPE (%)
Training	MLP ensemble	659.13	645.63	10.20	9.40
Testing		720.92	642.52	13.00	9.50
Training	MLP	628.28	569.67	9.80	7.60
Testing		700.29	780.71	11.10	8.50
Training	SVM	1464.34	1312.41	25.90	24.50
Testing		1396.86	1370.59	25.30	21.10
Training	Boosting tree	2012.51	1824.92	32.60	27.30
Testing		1736.34	1711.62	30.20	31.40
Training	Random forest	3061.98	2614.16	49.60	39.00
Testing		4328.04	3488.01	83.40	57.50
Training	Exhaustive CHAID	1471.81	1905.16	20.50	24.10
Testing		2236.16	2296.80	36.00	26.30

Figure 2.11: Performance prediction of several energy consumption models, source [28]

I wuold like to stress the mathematical formulation of Kusiak's optimization problem:

$$\min y_1(t+1) \tag{2.7}$$

subject to:

$$
\begin{aligned}
&y_1(t+1) = f_1(y_1(t-1), x_1(t), x_2(t), v_{11}(t), v_{12}(t), v_{11}(t), v_{13}(t), v_{11}(t-1), v_{14}(t), v_{15}(t), v_{11}(t-1), v_{16}(t)) \\
&y_2(t+1) = f_2(y_2(t-1), x_1(t), x_2(t), v_{21}(t), v_{22}(t), v_{21}(t-1), v_{22}(t-1), v_{23}(t), v_{23}(t-1), v_{24}(t), v_{25}(t)) \\
&y_3(t+1) = f_3(y_3(t-1), x_1(t), x_2(t), v_{31}(t), v_{32}(t), v_{32}(t-1), v_{31}(t-1), v_{33}(t-1), v_{34}(t)) \\
&50 \leq x_1(t) \leq 65 \\
&1.2 \leq x_2(t) \leq 1.8 \\
&49 \leq y_2(t+1) \leq 51 \\
&70.5 \leq y_3(t+1) \leq 71.5
\end{aligned}
\tag{2.8}
$$

where t-1, t and t+1 represent previous, current and next step. The variables $x_1(t)$, $x_2(t)$ are the configuration settings (supply air temperature and supply air static pressure set points), while v_i parameters are the uncontrollable variables such as infrared radiation, solar beam, wind velocity and direction. The constrained equation' system 2.8 can be turned into an unconstrained model:

$$
\begin{aligned}
&min(Obj1, Obj2, Obj3) \\
&Obj1 = y_1(t+1) \\
&Obj2 = max[0, 49 - y_2(t+1)] + max[0, y_2(t+1) - 51] \\
&Obj3 = max[0, 70.5 - y_3(t+1)] + max[0, y_3(t+1) - 71.5] \\
&50 \leq x_1(t) \leq 65 \\
&1.2 \leq x_2(t) \leq 1.8
\end{aligned}
\tag{2.9}
$$

Clearly the complexity and nonlinearity of the energy consumption does not allow to solve the equation system easily through traditional mathematical methods, hence some heuristich search methods have been implemented. In [26] to find near optimal solutions, Kusiak employed a classical Particle Swarm Optimization algorithm:

1. Particles are initialized randomly in terms of positions $x_i \in R^D$ and velocities $v_i \in R^D$ in the space.

2. Calculate the fitness for each particle.

3. The fitness of each particle is then put in comparison with the previous best values, if the current one is better than the previous, it substitutes $pbest_i$ with the current value and update the p_i and v_i of each particle. After that PSO set gbest to the best among all $pbest_i$.

4. Positions and velocities of the particles are revised using the following adjustment:

$$
\begin{aligned}
v_i &\leftarrow v_i + U(0, \phi_1) \cdot (p_i - x_i) + U(0, \phi_2) \cdot (p_g - x_i) \\
x_i &\leftarrow x_i + v_i
\end{aligned}
$$

5. Satisfied the stop criterion, p_g is the optimal solution and gbest is its fitness.

Refering to the i-th particle, (x_i, v_i) represent where it is positioned in the search space and where it is going to, while pbest is its best fitness reached so far. To stress $U(0, a)$ which is the uniform distribution in [0,a]. Recently several modifications has been proposals on PSO, in [27] Kusiak used three variants of it: CIWSPO, CPSO and DIWPSO. The key difference stands on the formula to update position and velocity of each particle, let's see each formulation:

- Constant inertia weight PSO (CIWSPO)

  $$v_i \leftarrow wv_i + U(0, \phi_1) \cdot (p_i - x_i) + U(0, \phi_2) \cdot (p_g - x_i)$$
  $$x_i \leftarrow x_i + v_i$$

  where w is an inertia weight that is constant.

- Constricted PSO (CPSO)

  $$v_i \leftarrow \chi(v_i + U(0, \phi_1) \cdot (p_i - x_i) + U(0, \phi_2) \cdot (p_g - x_i))$$
  $$x_i \leftarrow x_i + v_i$$

  where χ is the constriction coefficent, it controls the convergence of the particle.

- Decreasing inertia weight PSO (DIWPSO)

  $$v_i^{t+1} \leftarrow w^t v_i^t + U(0, \phi_1) \cdot (p_i^t - x_i^t) + U(0, \phi_2) \cdot (p_g^t - x_i^t)$$
  $$x_i^{t+1} \leftarrow x_i^{t+1} + v_i^{t+1}$$

  where w^t is a time function, described by the expression:

  $$w^t = \frac{w_{max}^t - t}{w_{max}^t} \cdot (w_{max}^t - w_{min}) + w_{min}$$

Another interesting heuristic algorithm used by Kusiak at [28] was a variant of Strenght Pareto Evolutionary Algorithm (SPEA) which, in particular, integrate a local search procedure (SPEA-LS) at each new generation in order to spot better solutions near each potential candidate. Let's describe briefly the algorithm:

1. Initialized a population P, create an external set, called PARETO for potential solutions.
2. Look for non-dominated solutions in P and store them in P*.
3. Calculate the fitness for each candidate in P and in PARETO sets.
4. Choose N_{parent} elements from the set P $\cup$ P* using tournament selection.
5. Extract randomly two elements among the N_{parent}, apply crossover and mutation and put the new individual obtained in a new set $P_{offsprings}$.
6. Apply a LOCAL SEARCH procedure to $P_{offspring}$. Each element is evaluated looking for n neighborhood solutions, the best one is chosen and compared with the current individual in $P_{offspring}$, if it is better, the new one replace it. P is then replaced by the set $P_{offspring}$. This local procedure has the purpose to improve the search.

7. Carry on till reaching the maximum number of generations.

An hybrid algorithm was the one employed at [29], called Strength multi-objective PSO (S-MOPSO), it is a moisture of SPEA and PSO, it tries to exploit the fact that SPEA, being an evolutionary algorithm, is suitable in looking for optimal global solutions, whereas PSO is optimal for searching for local solutions. In order to show how these optimization algorithms can really allow to save energy, Kusiak implemented S-MOPSO, getting the following results:

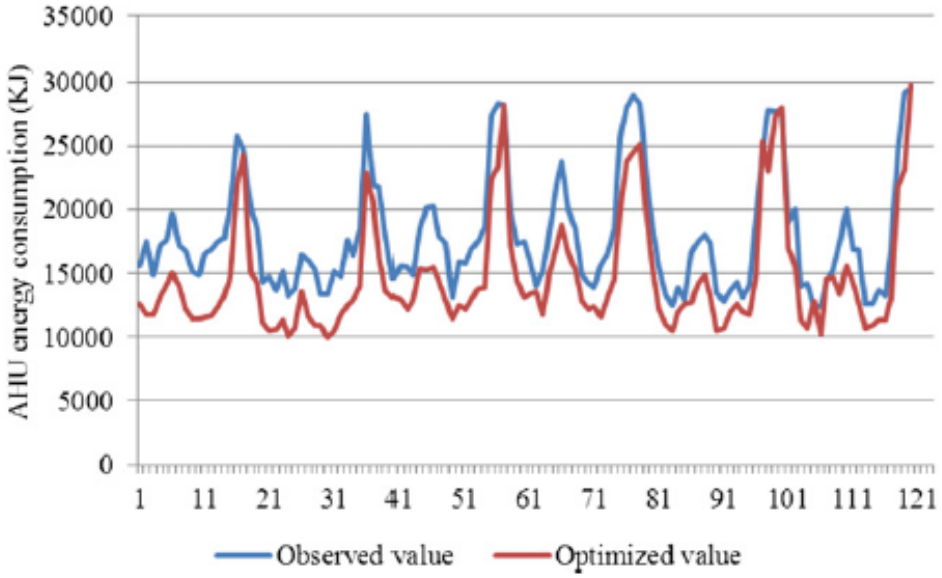


Figure 2.12: Observed and Optimized values for energy consumption, source from [29]

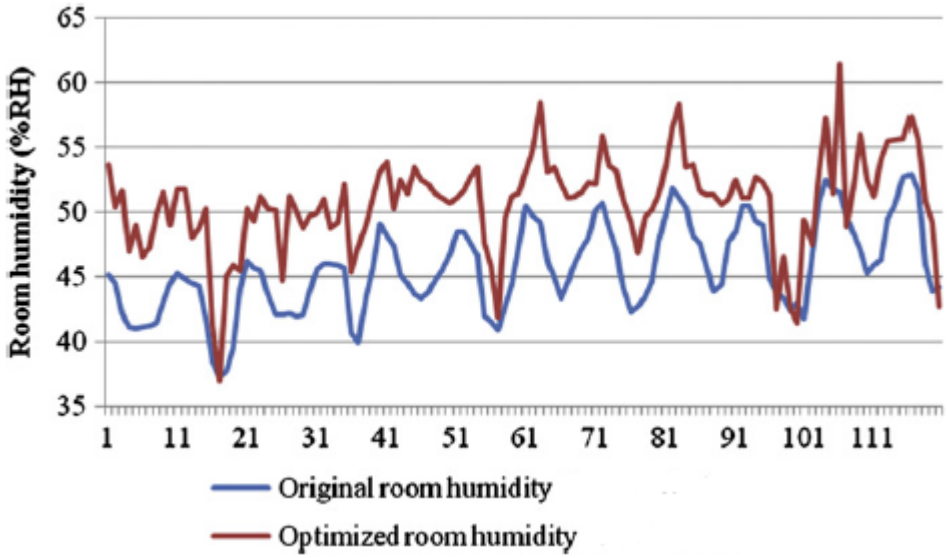


Figure 2.13: Observed and Optimized values for humidity, source from [29]

Figure 2.12 on the preceding page, shows us as choosing appropriate configurations of controllable variables, it is possible to maximizing the energy saving, in fact the optimized values (red line) is in average lower than the observed values (blue line). To what concern room humidity and temperature in figures 2.13 on the facing page and 2.14, they better adapt to the initial constraints as requested by the optimization problem.

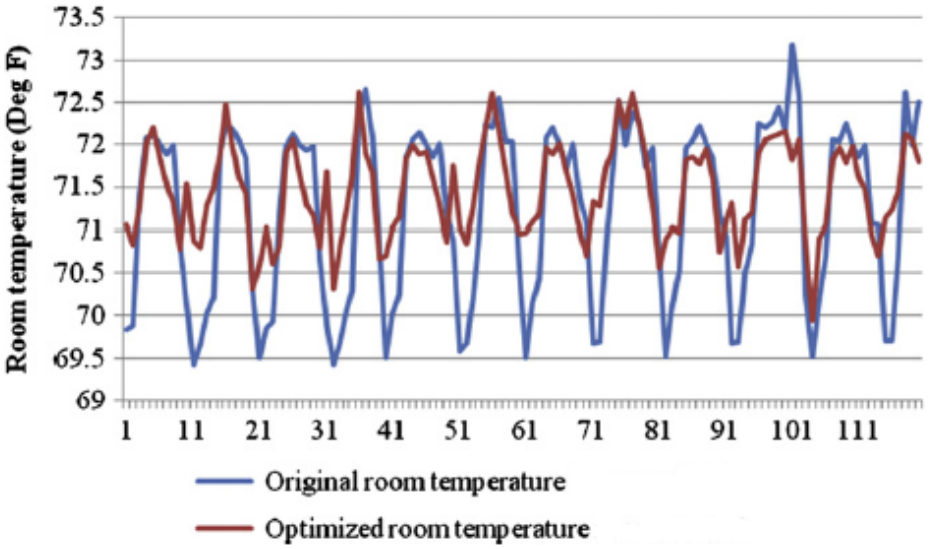


Figure 2.14: Observed and Optimized values for temperature, source from [29]

Chapter 3

Case of study: Stratega project

3.1 Introduction

Str.a.t.e.g.a project aims to reduce energy consumption of HVAC systems preserving optimal levels of comfort for occupants. A study case has been introduced in the first chapter, where three offices have been undertaken to sensory system for approximately three months. During this thesis I focused my attention on office 2, since it owns a completely automatic system that wuold allow my algorithms to have more control on it, in fact office 2 is the unique one which automatically set all the available controllable variables (d1, d2, b, fc). Hence, it is surely the perfect environment in which to elaborate my solution. The purpose is to find the best settings in order to reduce the future electric power (y) taking care of inhabitants comfort through PMV and DGI indexes (respectively x1 and x2). Specifically I wuold like to spot the best controllable variables combination of the future 6 hours so as to lower the energy consumption. Having abundantly described my ambition, let me illustrate, firstly, how this chapter is structured. In the very first section an explorative data analyses (EDA) is applied on my three datasets in order to interpret as much as possible the system's behaviour. Secondly I will implement a feature selection process so as to detect which variables influence more energy consumption, PMV and DGI. In the third section I will model my three objectives with different modeling approaches: Neural networks (NN), ARIMAX and random Forest (RF). The smallest section is the fourth one even if it is not surely the less important since a model validation is implemented. The best models will be then used to predict the future behaviour of the system. Finally an optimization algorithm is employed to find the best settings combination.

3.2 Exploratory data analysis

Exploratory data analysis defines an approach to analyse numerically and graphically a set of data before performing inference on it. The term was firstly used by Tukey [31] to inspire statisticians to explore data since from EDA is possible to spot outliers, see patterns, set up hypotheses and check assumptions. Several tools can be used in EDA, just to mention a few of them: numerically speaking there exist max, min, mean, median values for descriptive analysis, correlation to to understand the relationships among variables. To what concern graphical methods the most known are histograms, density distributions, and boxplots. However, a decision

maker has to choose the most appropriate for their purpose. Concerning my problem, after an initial descriptive analysis I identified as the most appropriate instruments:

- density distribution plot
- time series
- boxplot

3.2.1 Descriptive analysis

The following analysis shows some statistics that identify the general characteristics of a dataset. My dataset is composed by:

- qualitative variables: d1, d2, b, fc, v8, v13, v14
- quantitative variables all the others.

To what concern the first class I will show for each variable the frequancy distribution, while minimum, maximum, mean, median, first and third quartiles for quantitative variables.

```
d1        d2          b         fc              v1              v2
0:23479   0:21939   0  : 9040   0:24978   Min.   :17.09   Min.   :10.36
1: 5944   1: 7484   50 :   40   1: 4445   1st Qu.:22.27   1st Qu.:32.94
                    100:20343             Median :23.51   Median :41.13
                                          Mean   :23.03   Mean   :39.70
                                          3rd Qu.:24.61   3rd Qu.:48.30
                                          Max.   :26.75   Max.   :65.02
```

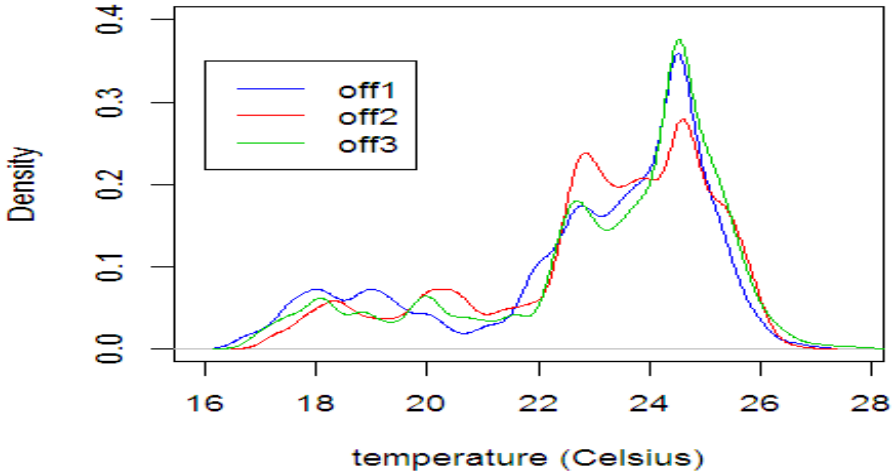
```
      v3          v4              v5               v6              v7
Min.   :0   Min.   :15.94   Min.   : 16.26   Min.   : 14.62   Min.   : 321.8
1st Qu.:0   1st Qu.:21.49   1st Qu.: 17.45   1st Qu.: 15.81   1st Qu.: 383.4
Median :0   Median :22.76   Median : 17.75   Median : 16.26   Median : 411.8
Mean   :0   Mean   :22.25   Mean   : 22.67   Mean   : 19.77   Mean   : 430.6
3rd Qu.:0   3rd Qu.:23.89   3rd Qu.: 23.71   3rd Qu.: 21.32   3rd Qu.: 455.6
Max.   :0   Max.   :26.22   Max.   :156.17   Max.   :215.47   Max.   :1424.0
```

```
v8              v9                 v10                 v11
0:14925   Min.   :-10000000   Min.   :-10000000   Min.   :  0.60
1:14498   1st Qu.:-10000000   1st Qu.:-10000000   1st Qu.:  2.60
          Median :-10000000   Median :-10000000   Median :  4.80
          Mean   : -6408246   Mean   : -6408252   Mean   : 45.77
          3rd Qu.:       18   3rd Qu.:        0   3rd Qu.: 38.20
          Max.   :       18   Max.   :        0   Max.   :741.20
```

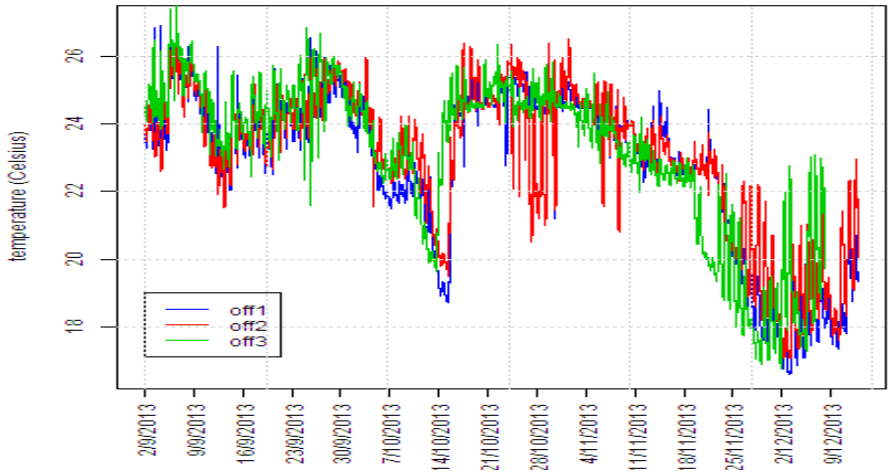
```
      v12             v13        v14             v15
 Min.   :-10000000   0: 8772   0:28019   Min.   :17.02
 1st Qu.:-10000000   1:20651   1: 1404   1st Qu.:22.29
 Median :-10000000                       Median :23.56
 Mean   : -6408230                       Mean   :22.97
 3rd Qu.:       61                       3rd Qu.:24.54
 Max.   :       61                       Max.   :27.38

       x1                 x2                y_load             ycum
 Min.   :-1.65816   Min.   :-49.3009   Min.   :-0.4493   Min.   :  0.00
 1st Qu.:-0.21377   1st Qu.: -0.8440   1st Qu.: 0.1040   1st Qu.: 15.90
 Median : 0.12605   Median : -0.5607   Median : 0.1393   Median : 60.00
 Mean   :-0.01603   Mean   :  2.6284   Mean   : 0.2879   Mean   : 63.02
 3rd Qu.: 0.41553   3rd Qu.:  2.9227   3rd Qu.: 0.2570   3rd Qu.:105.60
 Max.   : 1.04988   Max.   : 33.0361   Max.   : 2.3134   Max.   :146.60
```

At first glance we can realize that outside temperature (v9), illuminance (v10) and humidity (v12) present some problems since they obtain intolerable values, in fact the external station did not work well as confirmed by the company which collected the data. As a result these variables has been removed from our dataset. Another variable that has been removed is v3 that identifies the air velocity, which was expected to get values close to zero but not all the observations equal to zero. Finally I recall you that my response variable ycum (or simply y for the rest of the chapter) is cumulative, so it has to be treated as cumulative.

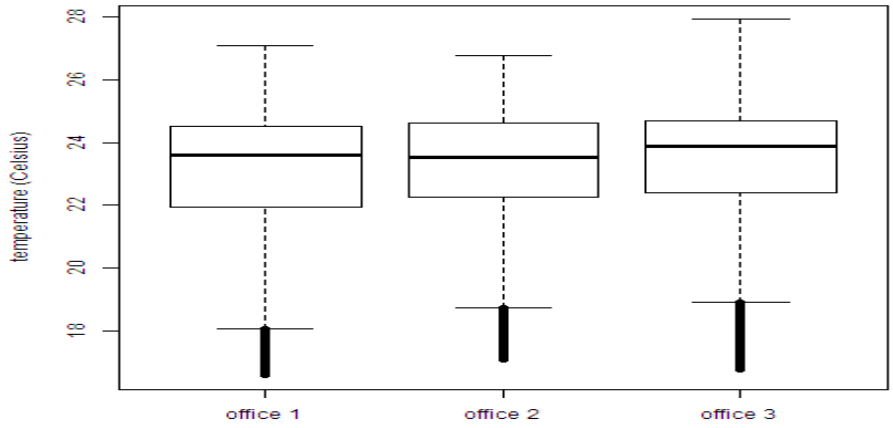
3.2.2 Graphical analysis



(a)

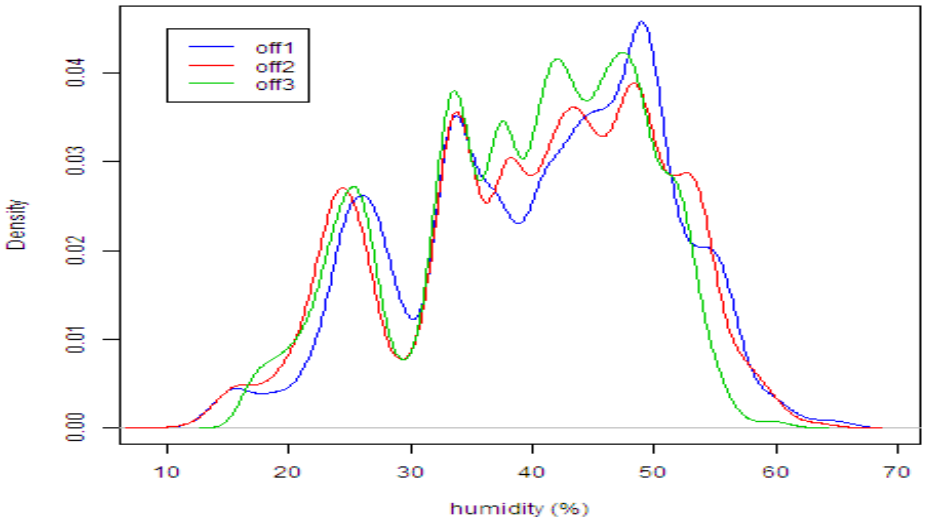


(b)

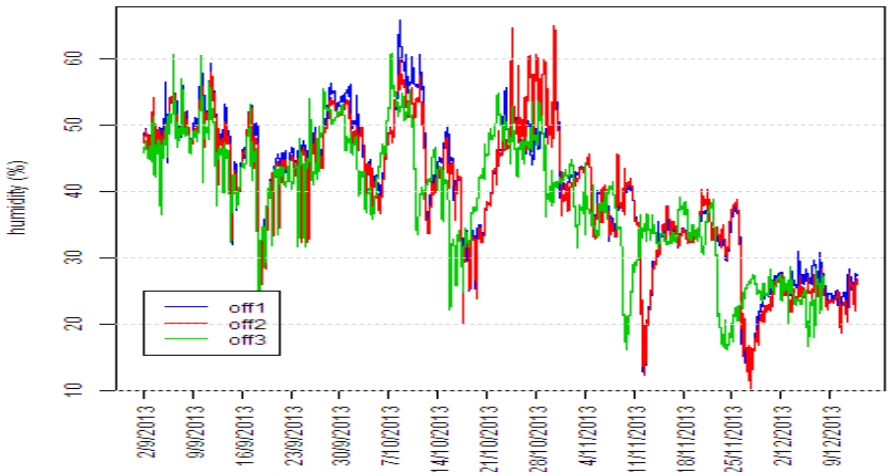


(c)

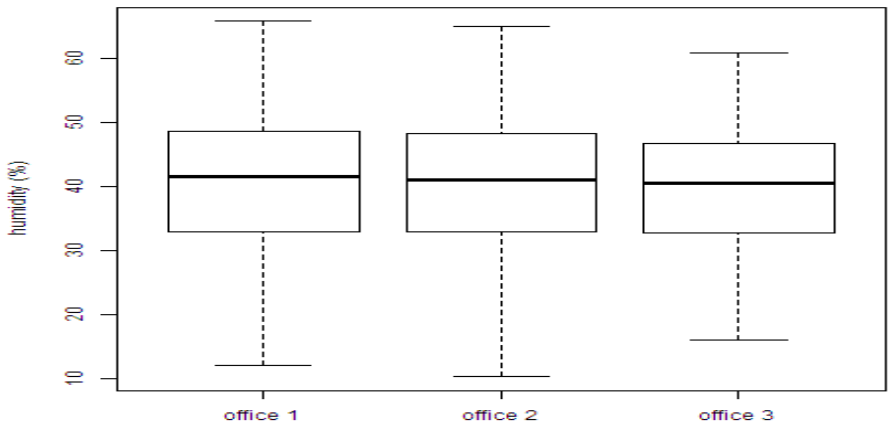
Figure 3.1: Density distribution (a), time series (b) and boxplot (c) of room temperature (v1) in offices 1,2,3 from 2nd September to 12th December 2013



(a)

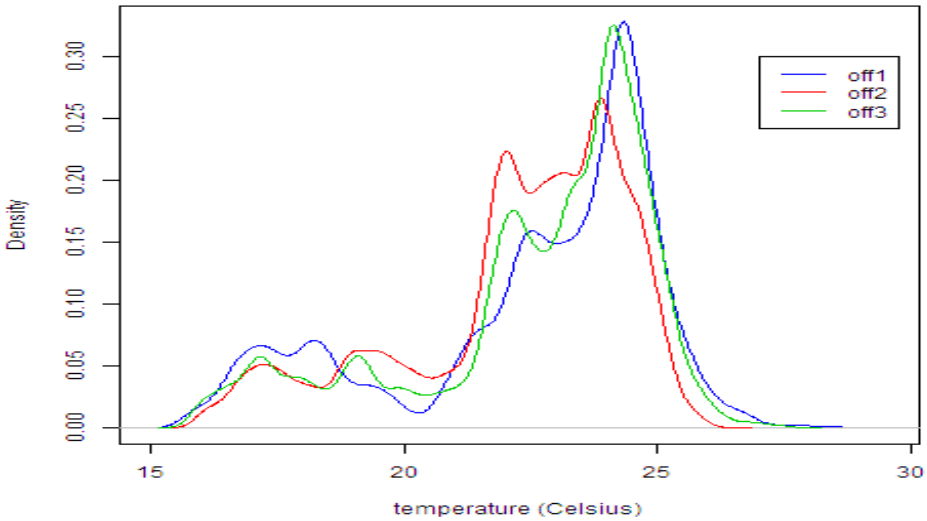


(b)

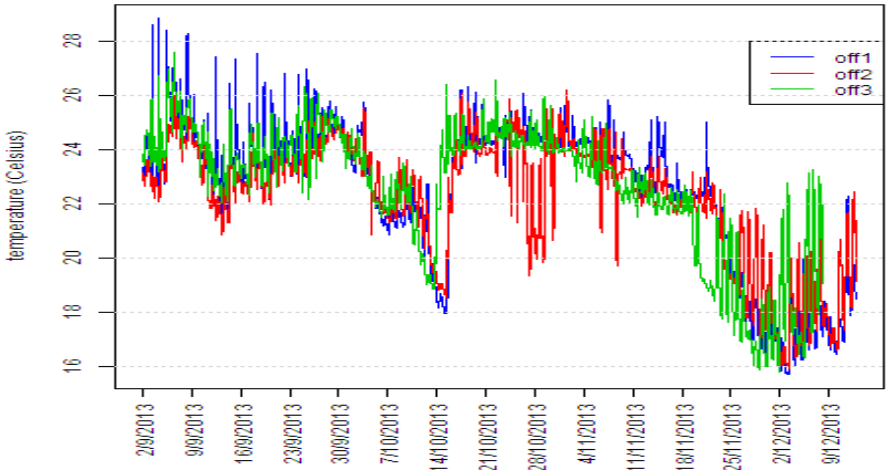


(c)

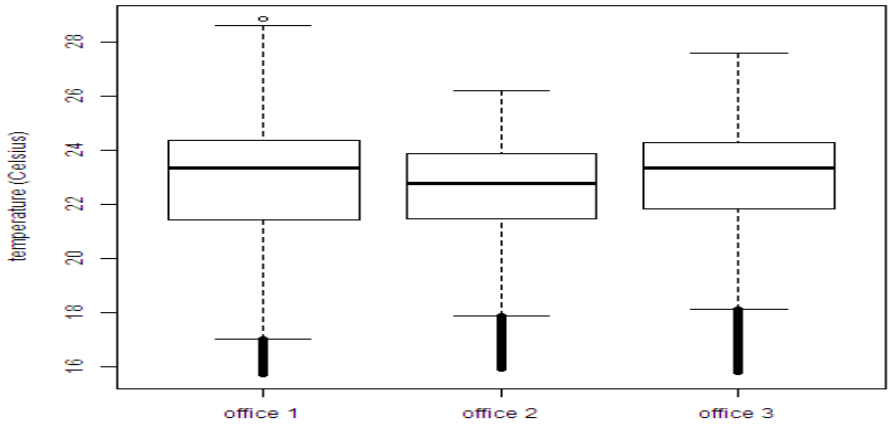
Figure 3.2: Density distribution (a), time series (b) and boxplot (c) of room humidity (v2) in offices 1,2,3 from 2nd September to 12th December 2013



(a)

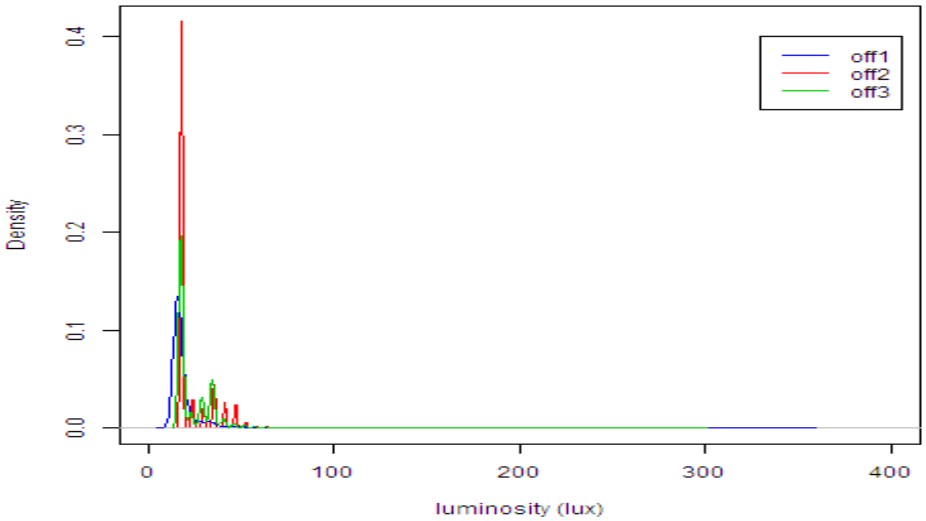


(b)

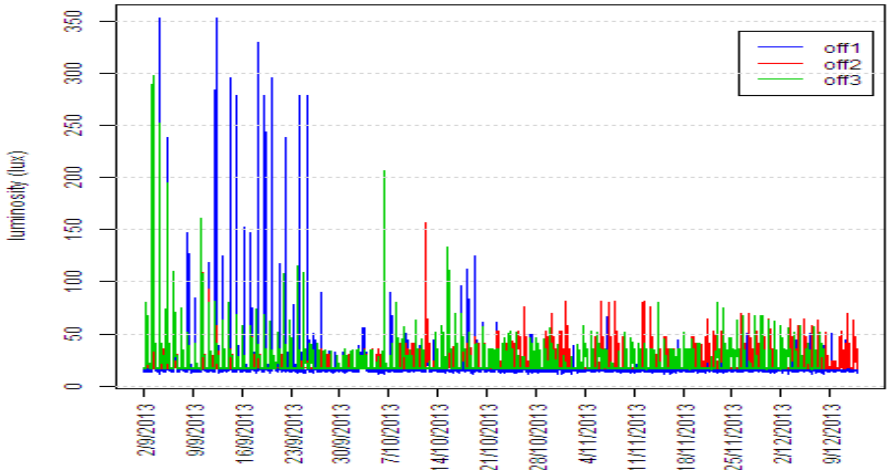


(c)

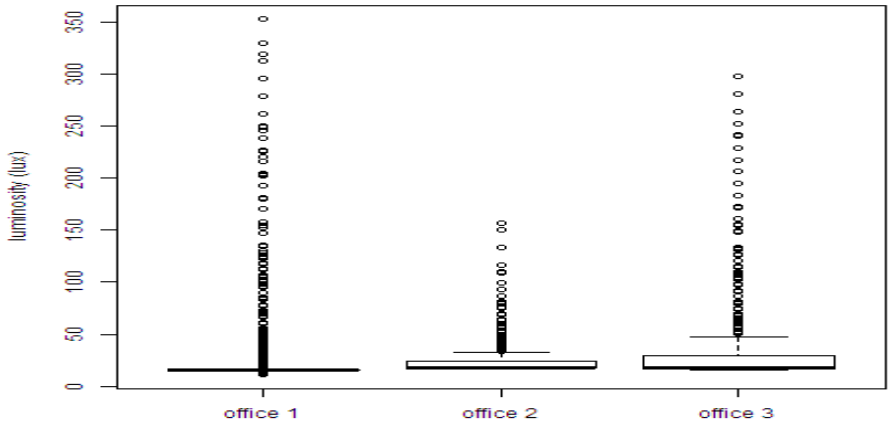
Figure 3.3: Density distribution (a), time series (b) and boxplot (c) of mean radiant temperature (v4) in offices 1,2,3 from 2nd September to 12th December 2013



(a)

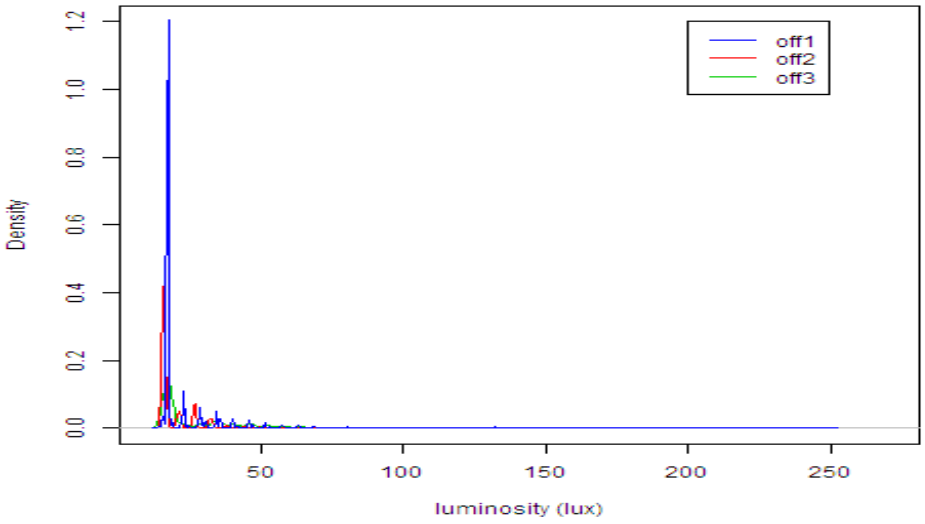


(b)

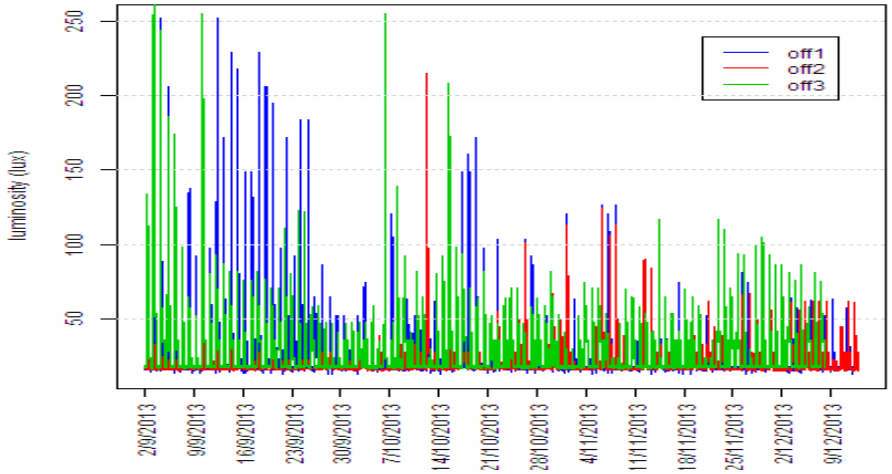


(c)

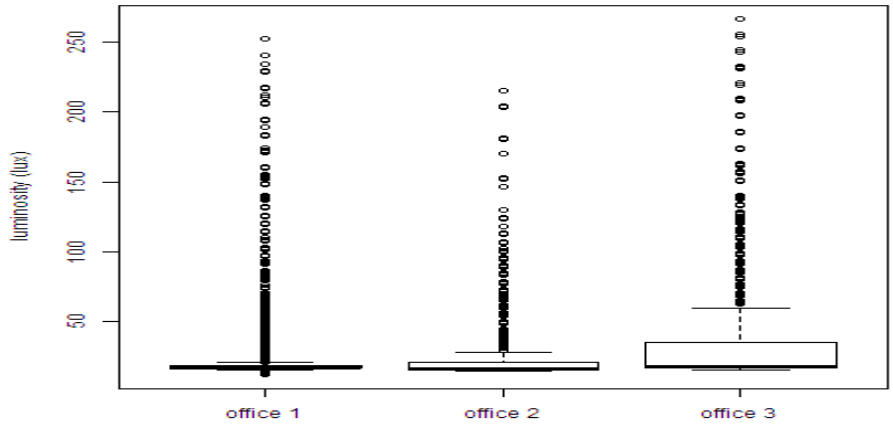
Figure 3.4: Density distribution (a), time series (b) and boxplot (c) of luminosity WEST (v5) in offices 1,2,3 from 2nd September to 12th December 2013



(a)

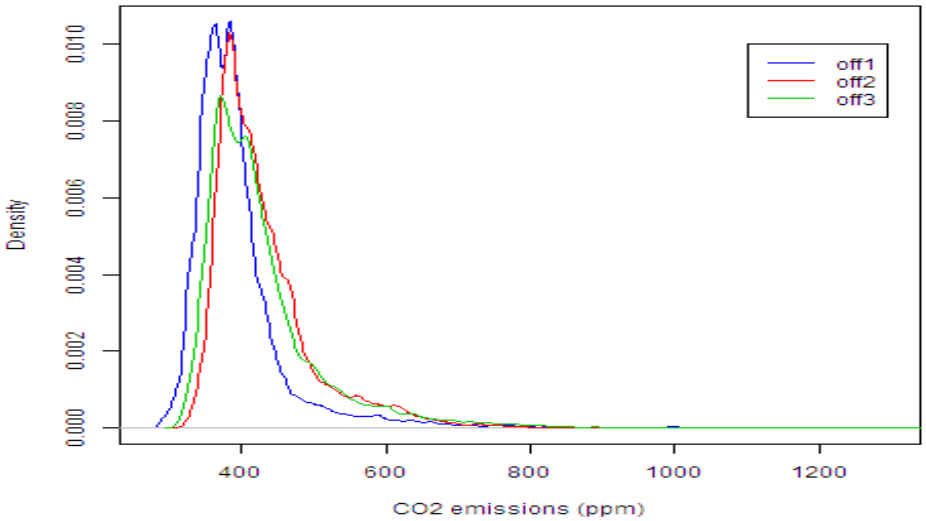


(b)

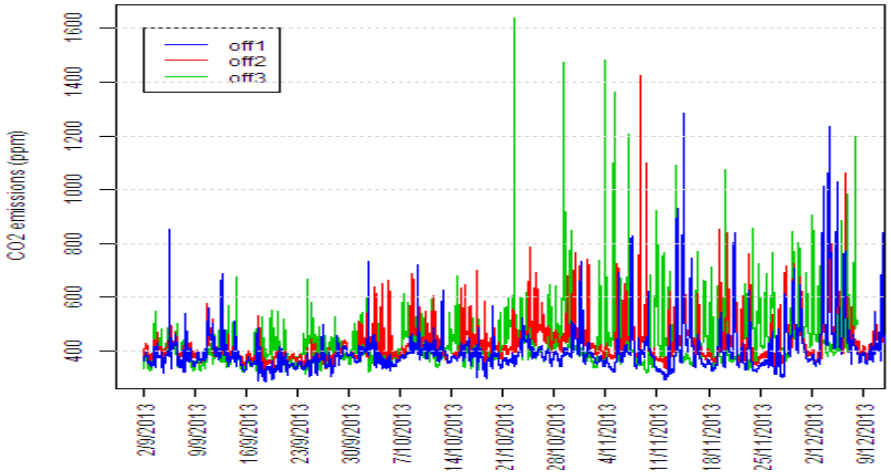


(c)

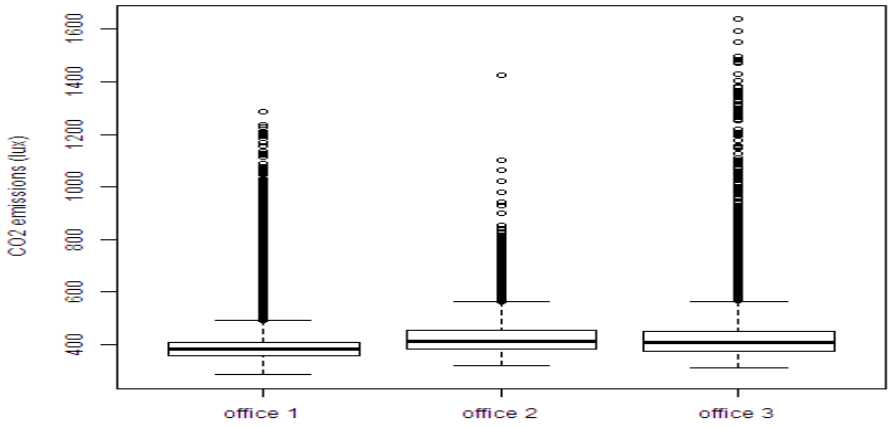
Figure 3.5: Density distribution (a), time series (b) and boxplot (c) of luminosity EAST (v6) in offices 1,2,3 from 2nd September to 12th December 2013



(a)

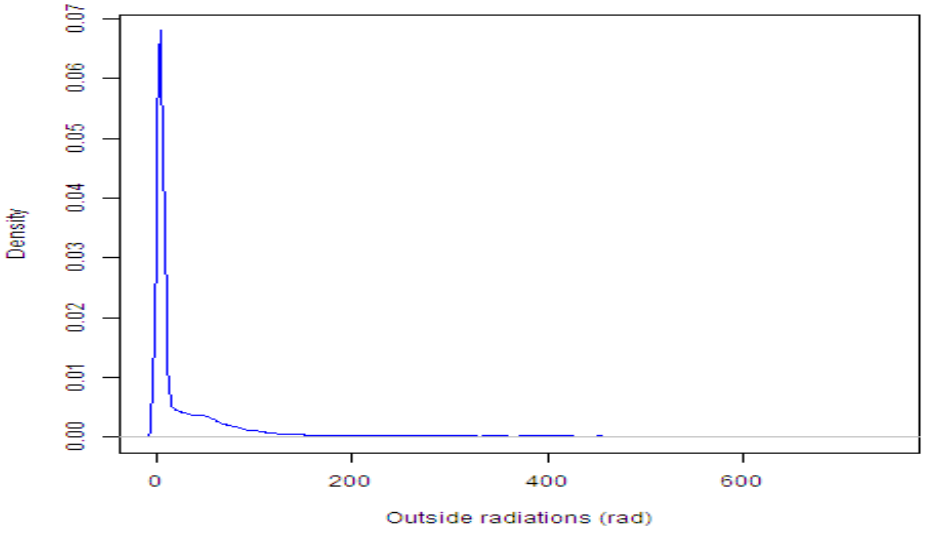


(b)

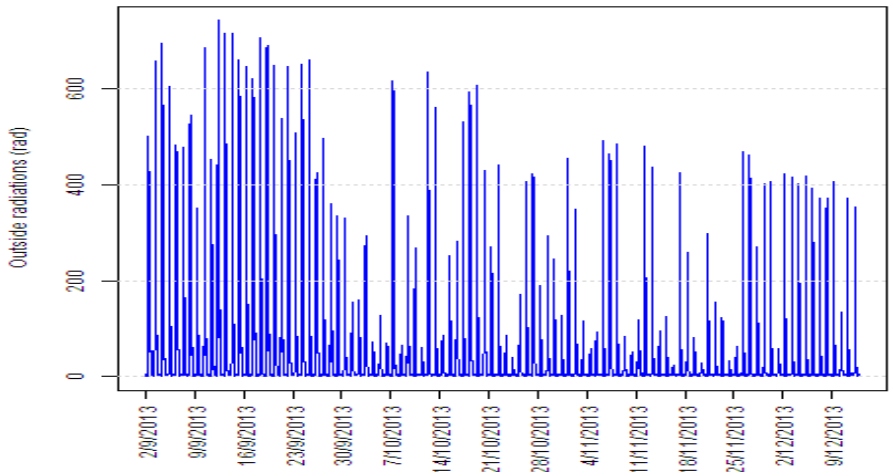


(c)

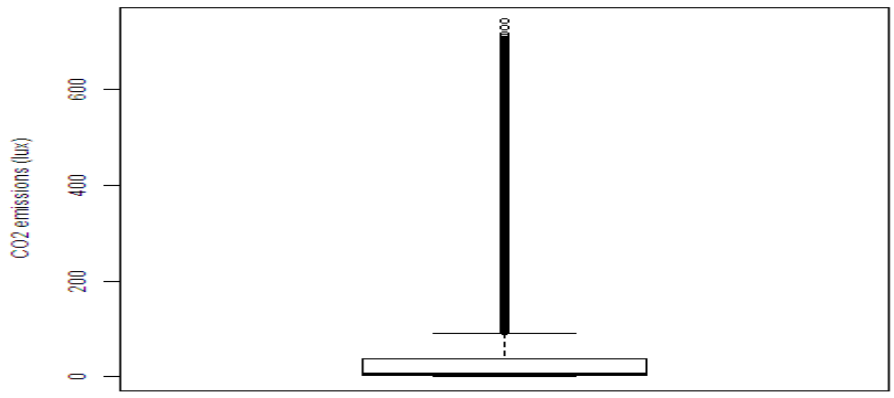
Figure 3.6: Density distribution (a), time series (b) and boxplot (c) of CO_2 presence (v7) in offices 1,2,3 from 2nd September to 12th December 2013



(a)

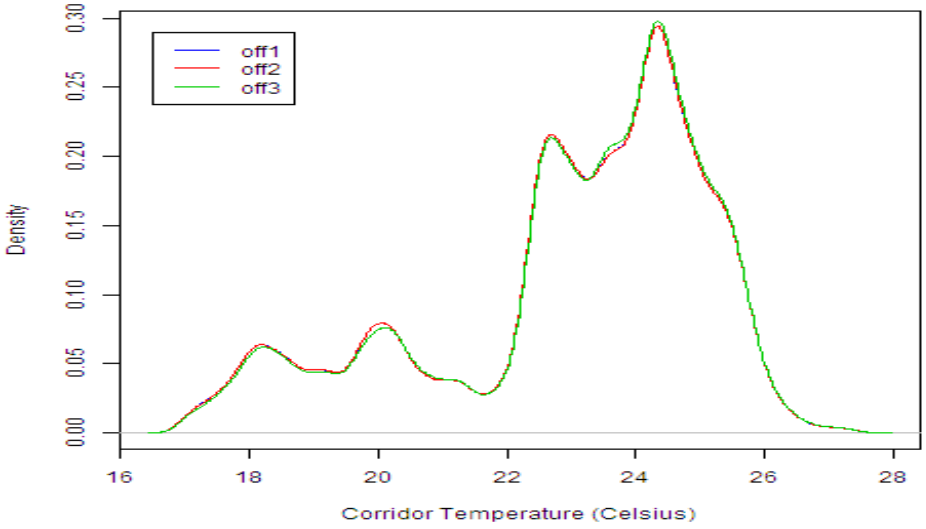


(b)

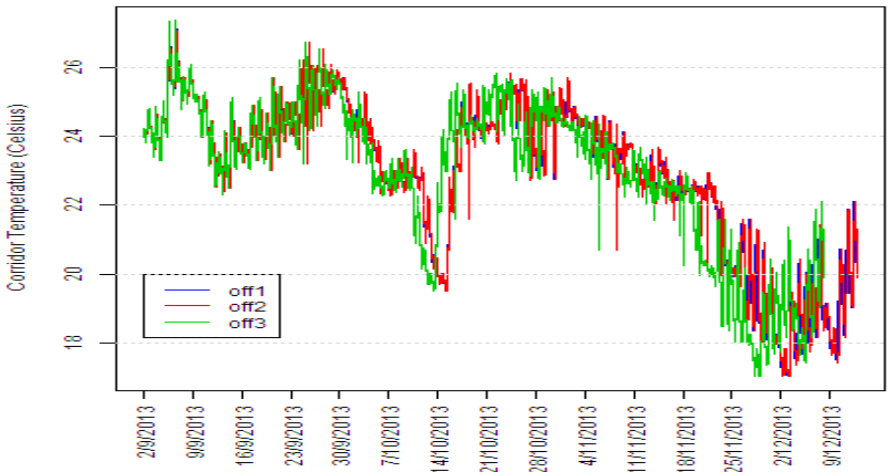


(c)

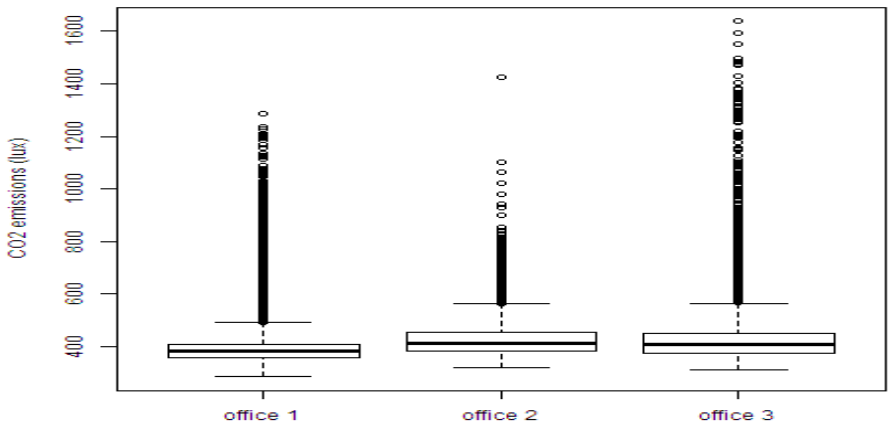
Figure 3.7: Density distribution (a), time series (b) and boxplot (c) of outside radiations (v11) from 2nd September to 12th December 2013



(a)

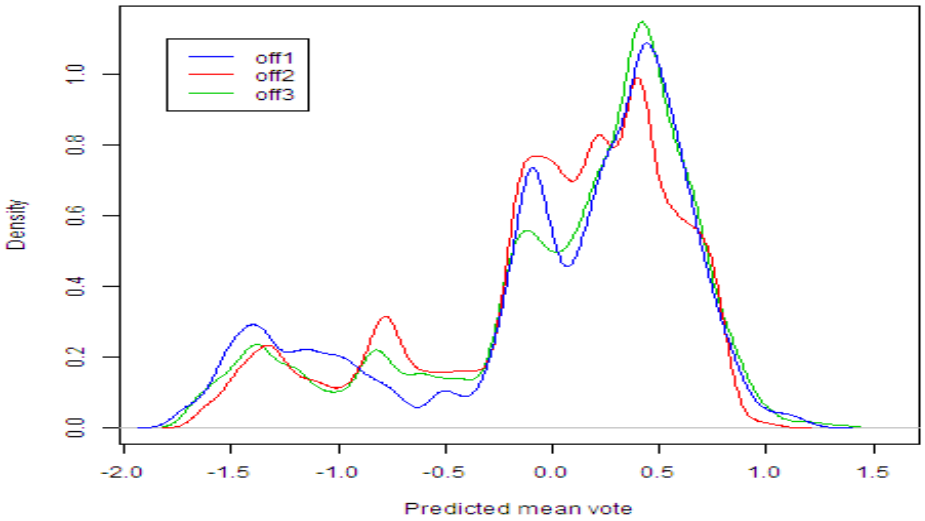


(b)

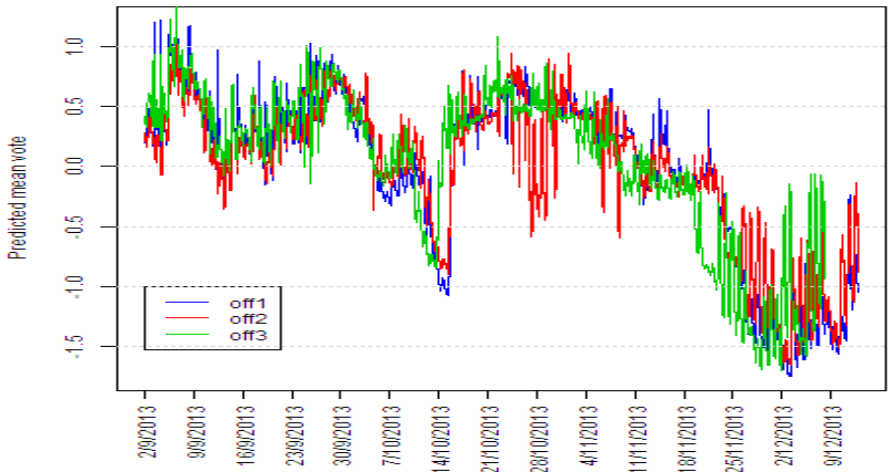


(c)

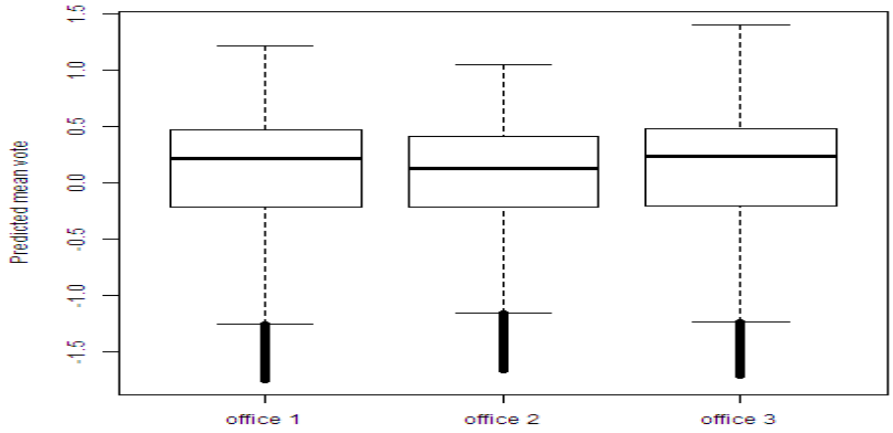
Figure 3.8: Density distribution (a), time series (b) and boxplot (c) of corridor temperature (v15) from 2nd September to 12th December 2013



(a)

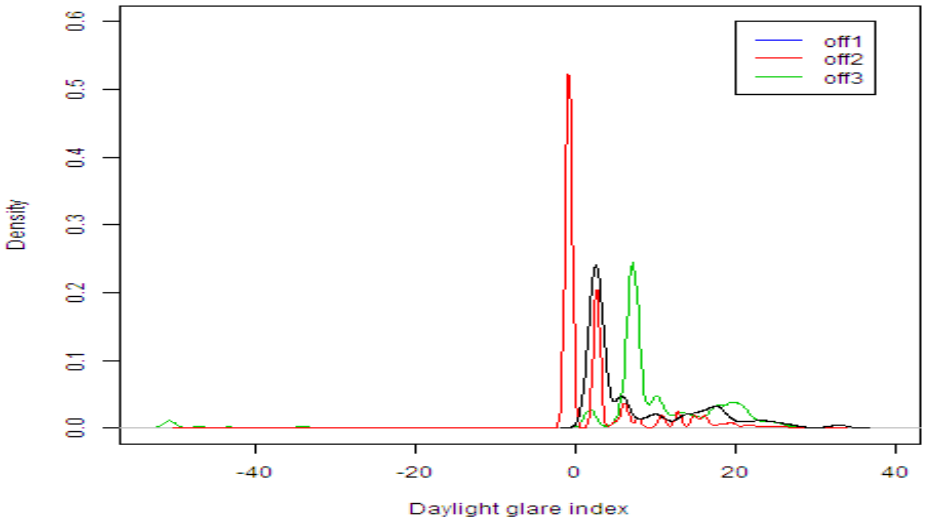


(b)

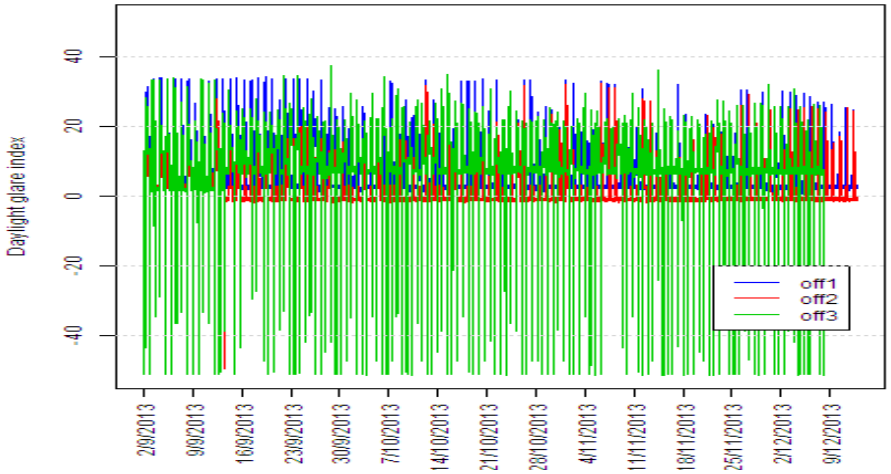


(c)

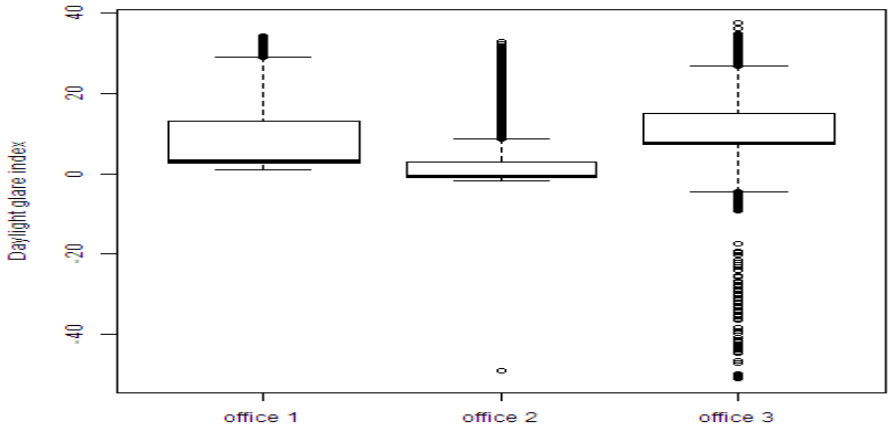
Figure 3.9: Density distribution (a), time series (b) and boxplot (c) of PMV (x1) in offices 1,2,3 from 2nd September to 12th December 2013



(a)

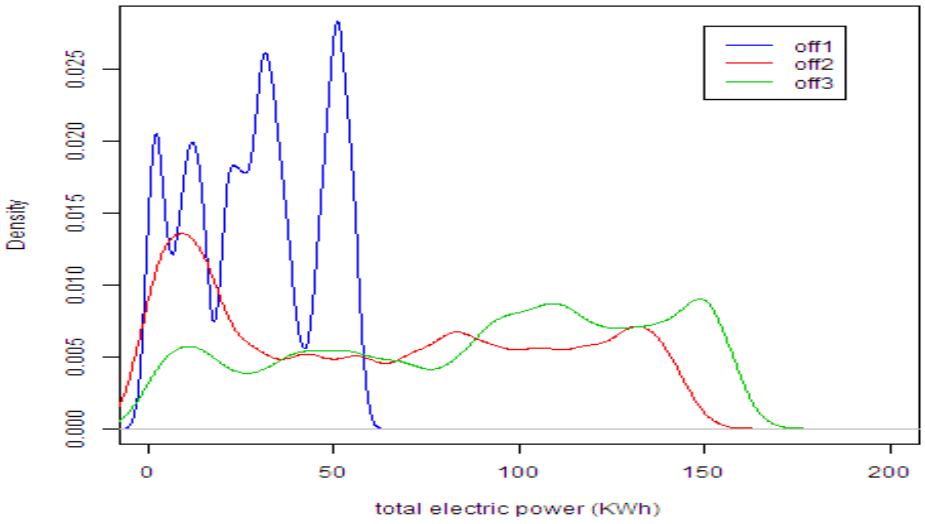


(b)

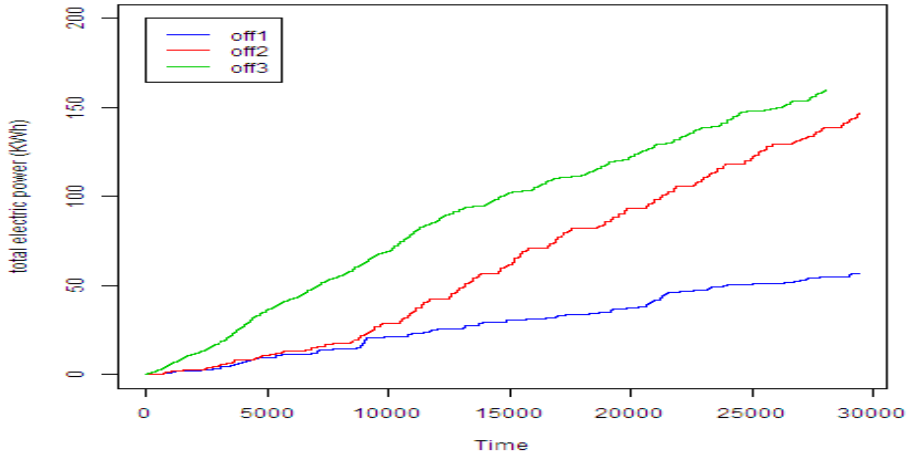


(c)

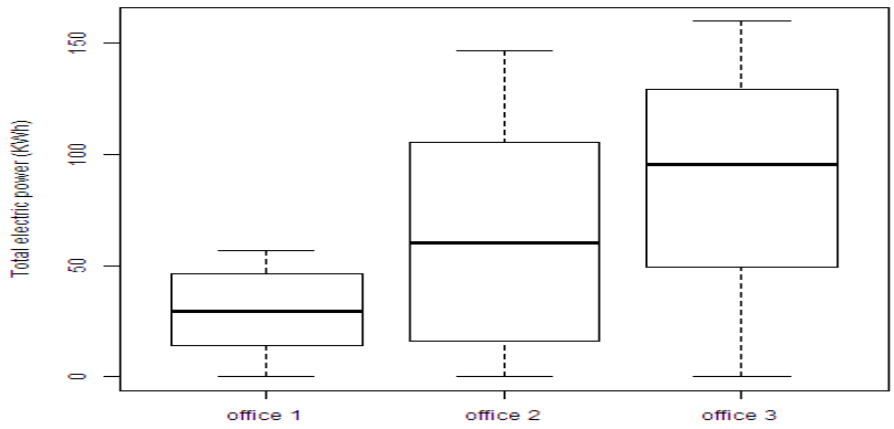
Figure 3.10: Density distribution (a), time series (b) and boxplot (c) of DGI (x2) in offices 1,2,3 from 2nd September to 12th December 2013



(a)



(b)



(c)

Figure 3.11: Density distribution (a), time series (b) and boxplot (c) of total electric power (y) in offices 1,2,3 from 2nd September to 12th December 2013

The internal temperature (v1) is similarly distributed in all three offices with median around 24℃and presents the same behaviour for all the offices. Their distributions (see 3.1 on page 51) exhibit a longer left tail (visible in figure 3.1 on page 51) due to the presence of outliers (evident in the boxplots) that can be temporarily speaking individuated in two periods from 7th to 14th October and from 20th November to the end of the detection period. Also humidity (v2) is similarly distributed in in all three offices (figure 3.2 on page 52), the medians are slightly greater than 40% of humidity, however a big variability is shown in figure 3.2 on page 52 (c) since the min and the max values are respectively at 10% and 70%, as a result PMV (x1) will be surely affected by this strong variability. Looking at the time series plot (figure 3.2 on page 52 (b)) a decreasing trend can be easily observed, but it is quite normal since in winter season a lower humidity is expected. Regarding the mean radiant temperature (v4) is expected just a slight difference with respect to internal temperature, in fact the medians are 1℃lower than v1 for all three offices. Besides v1 and v4 present a similar trend over the time and also the peaks of minimum are situated in the same time points. To note that in average office 2 obtains lower values in terms of room and radiant temperature.
To what concern EAST and WEST luminosity in the room (respectively v6 and v5), high peaks are evident daily in the morning (figures 3.5 on page 55 and 3.4 on page 54 (b)) because of the sunlight since the three rooms are faced toward EAST. However, office 1 and 3 exhibit very strong peaks in September, this fact was explained me by the company which collected data: September was scarsely crowded by people in the building, thus the windows in that period were purposely open in office 1 and 3. The boxplots show these peaks as outliers.
Another interesting variable is v7 which describes the presence of CO2. Nevertheless, it can help me to spot also the presence of people. For instance, in September the low presence of people in the building is described by low values of CO2 (see figures 3.6 on page 56 (b)). On the other hand very high peaks of CO2 are visible from the 21st October 2013, thus I expect that an unusual occupancy was verified in that period such as meetings.
As shown by the boxplot and the density distribution skewness characterizes the outside radiations, vast numbers of ouliers are due to the high peaks of radiations that reach the building especially in the morning.
As demonstrated in chapter 2 PMV depends on internal and radiant temperature, then is not surprising whether the PMV trend follows a very similar pattern over the time to the ones of v1 and v4. The density distribution (3.9 on page 59 (a)) shows a long left tail due to the temperature people sensation in December, which is regularly out of the wellness limit condition ($-0.5 \leq PMV \leq 0.5$).
Concerning the daylight glare index what immediately capture my attention looking at 3.11 on the previous page (b) is the high number of DGI mimimum peaks shown only by office 3, which immediately let me think that a problem occurs on the DGI sensor. Unfortunately sometimes the DGI sensor gets blocked and returned those minimum values. As a result in figure 1 I decided to cut off the left part in order to concentrate only on correct values.
The most compelling figure concerning the total electric power is definitely 3.11 on the preceding page (b). Office 1 and 2 seems to consume approximately the same energies till the first days of October, where the trend related to office 2 raises and get far from office 1 consumption. On the

contrary office 3 consumes a double energy with respect to the other two untill the beginning of October, where its trend changes consuming energy less than before.

3.3 Features Selection

Nowadays when you deal with modern dataset, you deal with a large and complex amount of data, however what actually it could be considered a problem is the high dimensionality that today's datasets present. For instance, in bioinformatic field, datasets with thousands of features (or variables) is become an expected characteristic. Hence, variable selection is a fundamental step in modeling. It provides the least number of dimensions, among all the features, that will supply my learning algorithm. Generally speaking, the selected variables own a couple of distinctive characteristics as relevance and irredundancy. The first property indicates the feature has an evident impact on the output (in my case either energy or PMV or DGI), while the second one explicits that no other variable can replace it (no one has the same effects on the output). The reasons why selection feature is so important can be easily explained: first of all, it reduces (hopefully removes) noyse data, then raise the data quality, speed up algorithms that will be applied in future and finally it can simplify and improve the accuracy of the models. The methods used by feature subset selection can be branched in:

- Filter approach: it is performed before the modeling process and it is independent from the learning algorithm applied.
- Wrapper method: here, the learning machine is directly used to select variables, the process needs an iterative performance feedback, to test the importance of a subset of variables.

The approach used for my problem is a filter method based on Random Forest (RF), additionally it is supported by a Correlation analysis among variables, the latter will confirm or discard the selection performed by RF. My choice is fallen on filter methods since they are less intensive computationally speaking, in fact the algorithm is performed once, and it will be not dependent on the machine learning that I will use in the next section for modeling. In subset selection "ensemble learning tree" is a very common choice, the term "ensemble" indicates that it generates many weak learners (in this case many trees) and it ends up assembling all their results. Among those, we have Boosting Tree [32], Bagging [33] and the most recent one Random Forest introduced in 2001 by Breiman [34]. Random Forest can be further used for regression and classification and its robustness against overfitting problem yields it an optimal tool. Let's describe how Random forest works:

1. Divide the dataset in train set and test set, the latter set will be called "out of bag" from now.
2. Extract n bootstrap samples from the trainset
3. Construct binary regression tree for each bootstrap sample (or classification tree whether you are dealing with classification problems), and during the splitting phase instead of choosing the best split node among all the p variables (as in Bagging procedure), it determines the best split variable choosing among m variables randomly selected ($m < p$).

4. The prediction is then carried out gathering the prediction of each tree, in terms of regression it uses the simple average among the n tree's results, while in classification problems majority vote is used.

The parameters to set accurately in Random Forest are principally: the number of binary trees to generate, the number of variables randomly selected as candidates at each splitting phase and the depth of the tree. An important role, to what concern prediction, is computed by the "out of bag" data (OOB) which is the data excluded by the bootstrap process at the beginning (step 1), in fact using a prediction error metric this exclusion allows to see how much the model is accurate.
Coming back to our principal topic (subset selection), the importance of each variable is determined simply permuting permissive values of that variable (leaving unchanged the others) in the OOB data, then calculating whether the prediction error increases. Undoubtedly whether the prediction error increases, the variable will be more relevant. In my case, the Mean Square Error (MSE) has been used as prediction error even if different metrics are acceptable:

$$MSE_{OOB} = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y_i^{OOB}})^2 \tag{3.1}$$

where $\hat{y_i^{OOB}}$ is the average prediction of the i-th observation on the all trees. If you want to rank in order of importance you can add the following formulation:

$$IncMSE_{OOB} = 1 - \frac{MSE_{OOB}}{\hat{\sigma_y^2}} \tag{3.2}$$

where $\hat{\sigma_y^2}$ is the average variance (calculated with n divisors). Regarding feature selection of my case study, some considerations need to be done. Since my final purpose is to find the best configuration of the controllable variables (d1, d2, b, fc), they will be surely present as explanatory variables of each model, therefore, I do not need to implement them in the features selection process. Moreover some categorical variables have been purposely removed from our final models such as v8, v13, v14. This arguably choice has been taken since their effects can be detect paying more attention on some uncontrollable variables, for example the occupancy (v8) could be observed by an increment on CO_2 emissions(v7). Explained some considerations, let's show the results obtained by feature selection via RF:

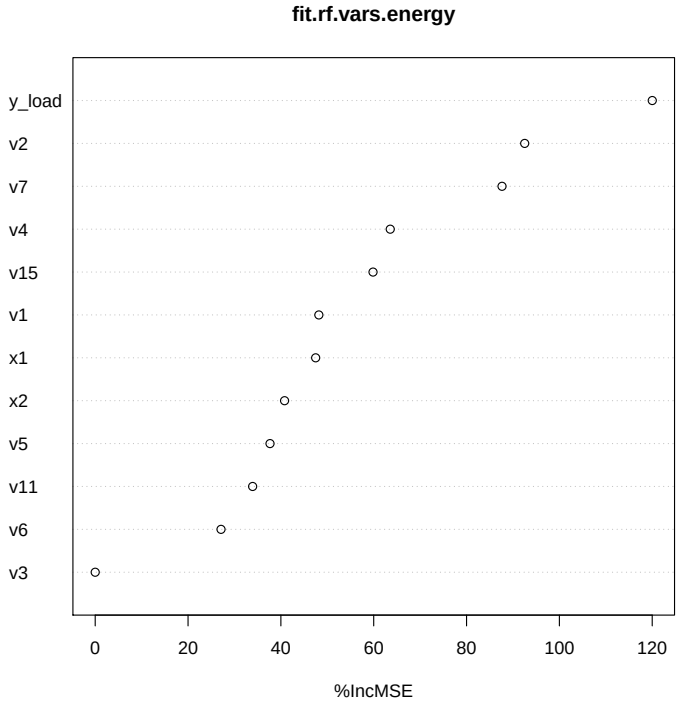


Figure 3.12: Variable importance for total electric power model

The results of the three models are represented in three different figures, in each one we find on the y axis the variables in order of importance, while in the x axis there is the IncMSE.
In my opinion simply removing variables through a threshold could be an option, however it wuold not be reasonable since valuable considerations have to be done. In terms of energy consumption are considered relevant by RF the following features: room temperature (v1) and humidity (v2), radiant mean temperature (v4) and certainly the electric power due to only heating (y_load) since the latter should be contained in the total electric power. What actually capture more my attention is the high score of CO_2 emissions (v7) which seems to influence the total electric power, speculating on it I could think that rising of CO_2 particles is symptom of occupancy in the room, then CO_2 warms the surrounding environment consequently relaxing the heater. Last but not least variable to join the model is the temperature of the outside corridor (v15), I could guess the door is often open, so being a corridor usually a bit colder than a room, the system needs more power to warm the entire environment (room and corridor).
Certainly predictors as room temperature and humidity, mean radiant temperature affect PMV, this can be proved simply nooting that they take place in the Fanger's equation (chapter 2). An impressive result has been obtained by the outside radiations (v11), then probably the good position of the building and the window allow radiations to get easily in (heat transfer for radiations). In addition, CO_2 emissions (v7), y_load and the EAST luminosity (v6) probably will be part of the PMV model.

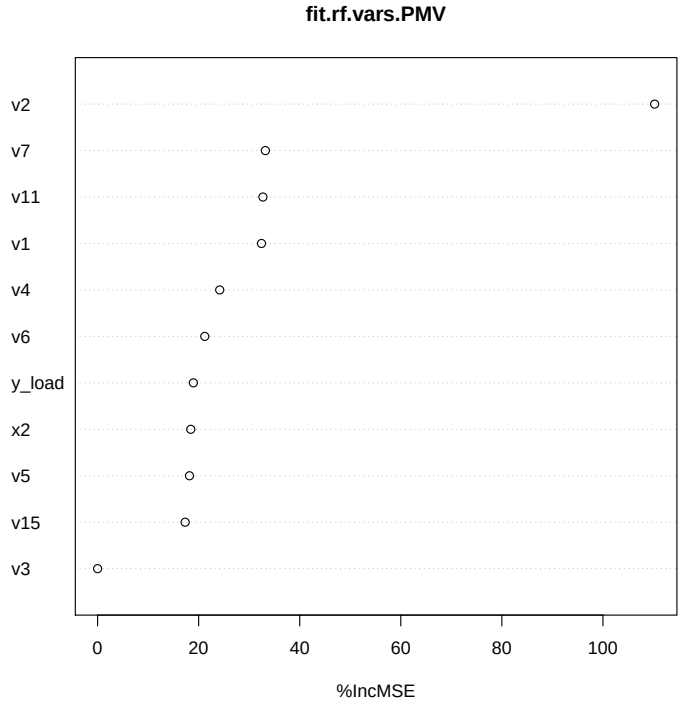


Figure 3.13: Variable importance for PMV model

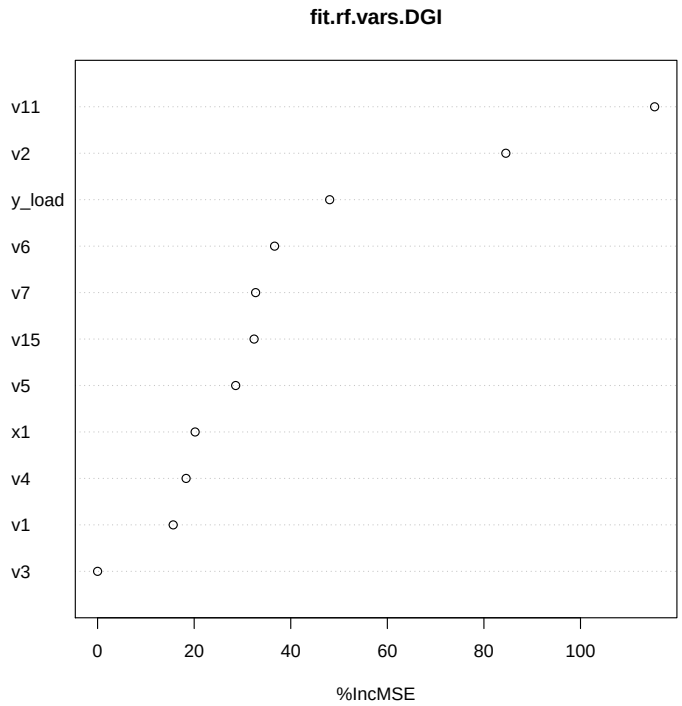


Figure 3.14: Variable importance for DGI model

As expected in DGI model will join the EAST and WEST luminosity (v5, v6) and the outside radiations (v11). What actually surprises me is the high scores obtained by internal humidity (v2), corridor temperature (v15), thermal load (y_load) and CO_2 emissions (v7), anyway they will be compared with the results obtained by the Correlation analysis to be part of the model. In order to verify the responses given by RF, a correlation analysis is applied among response variables and explanatory ones. Correlation is a term that can be easily explained using an example: today it is not surprising that longer height results in longer foot, the study [35] proved that height of Mumbai individuals is 6.5 times the lenght of their foot size, thus it means there is a relationship between these two lenghts. The latter example clearly describes a linear relationship, but more important is to know how strong is this correlation, to know it correlation coefficients have been introduced in history. Given two variables, let's say X and Y, the most known correlation coefficient was invented by Pearson:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \tag{3.3}$$

where $COV(X,Y)$ is the covariance between X and Y. The $\rho_{X,Y}$ coefficient can assume $-1 \leq \rho_{X,Y} \leq 1$, in particular:

- $\rho \approx 1$ indicates strong linear positive correlation
- $\rho \approx 0$ uncorrelation between X and Y
- $\rho \approx -1$ strong linear negative correlation.

The Pearson coefficient requires X,Y linear dependent and normally distributed. As on of the latter requirements are not satisfied another coefficient needs to be adopted such as Spearman coefficient. Differently from Person coefficient, it can handle with non-linear dependency which is what I need to solve my problem, however Spearman coefficient requires monotonicity relationship between X and Y. This property allows to detect esponential and polynomial relationships for instance, furthermore since linear relation incolves monotonicity through Spearman coefficient is also possible to detect linear relationship. In my study case having more than two variables to compare, a correlation matrix is necessary (figure 3.15). Empty cells represent uncorrelated variables, while blue and red coefficients are respectively positive and negative correlationships. The total electric power is negatively correlated with room temperature (v1) and humidity (v2), mean radiant temperature (v4), PMV (x1) and y_load. The predicted mean vote (x1) is positively correlated with room temperature(v1) and humidity (v2), mean radiant temperature (v4), and the corridor temperature (v15). Finally the daylight glare index measures a correlation with the outside radiations (v11). These correlations support the choices done by RF, in the next section our models will be formalized.

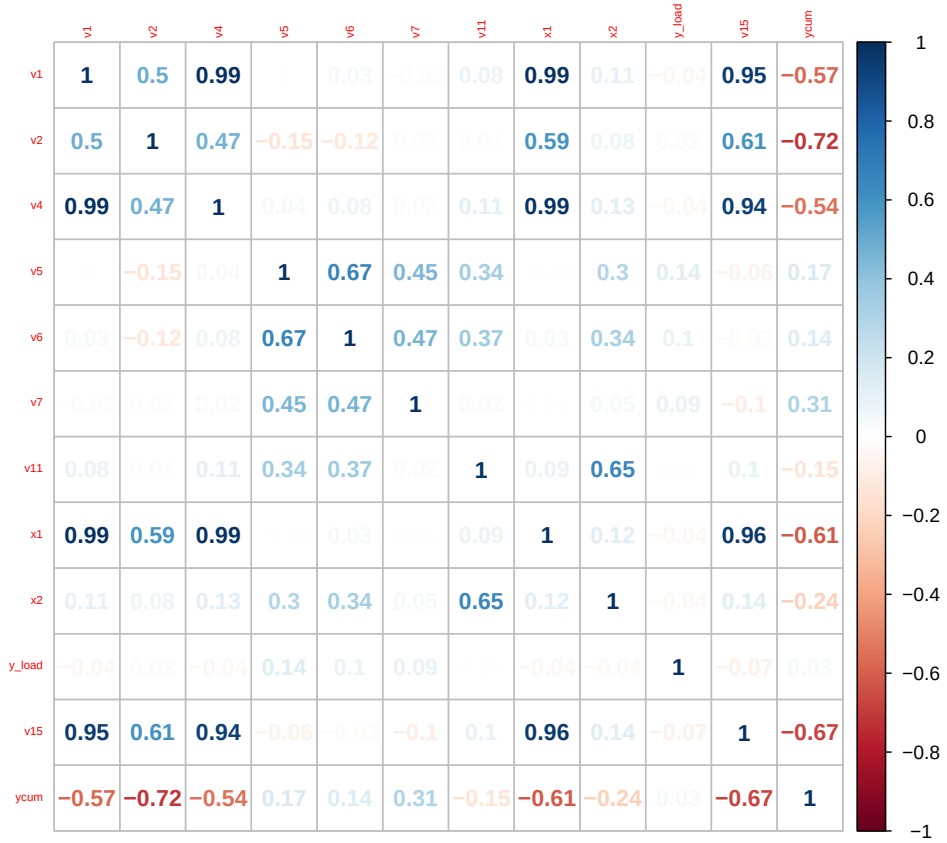


Figure 3.15: Spearman correlation matrix

3.4 Modeling

Very often, not only in academic field, we hear the term model used for the most diverse purposes but sometimes it is abused or inappropriate used, thus I wuold like to explain firstly its real meaning, in particular from a statistics point of view. A model is, metaphorically, a picture of a sytem or event that allows us to better understand how it works, in other words I could say that a model describes a certain system even though it does not describe it in each single detail. For instance, an architect builds models in everyday life in order to present, to show ideas and for sale purposes. Engineers are constantly in contact with models as to plan new aircrafts, ships and rockets. Therefore, modeling permits people to reproduce approximately real facts and systems. Statistically speaking, I will use models to explore, test and predict the behaviour of 3 components: total electric power (will be described by function y), predicted mean vote (x1) and daylight glare index (x2). To note that I am dealing with three continuos response variables, therefore I am facing a Regression problem. Using those features selected in the previous section, I define mathematically the cost funtions in which I will work on:

$$y = f(d1, d2, b, fc, v1, v2, v4, v15, y_{load}, x1, x2) \tag{3.4}$$

$$x1 = f(d1, d2, b, fc, v1, v2, v4, v7, v11, y_{load}) \tag{3.5}$$

$$x2 = f(d1, d2, b, fc, v2, v5, v6, v11, y_{load}) \tag{3.6}$$

As you can see, my intent is not to show mathematically how much each variable influences the responses, but to define precisely which variables affect them.
Likewise Kusiak 's work [27], I decided to extent my three models adding also the previous lag of the response variable itself in order to take into account the previous state of the model, now the cost functions take the following shapes:

$$y = f(d1, d2, b, fc, v1, v2, v4, v15, y_{load}, x1, x2, lag(ycum, 1)) \tag{3.7}$$

$$x1 = f(d1, d2, b, fc, v1, v2, v4, v7, v11, y_{load}, lag(x1, 1)) \tag{3.8}$$

$$x2 = f(d1, d2, b, fc, v2, v5, v6, v11, y_{load}, lag(x2, 1)) \tag{3.9}$$

The reader is probably asking the reason behind the choice to add lagged dependent variables, it is because those lagged variables will provide me even more precise parameters estimation, as a result my models will improve accuracy during future prediction phases. Those three formulations will be now modelled using three different learning machines:

- Artificial Neural Network (ANN or simply NN).
- ARIMAX.
- Random Forest (RF).

Firtsly, the dataset needs to be decompose in training and testset. Common use is to divide the dataset as follows:

- trainset: 65% of the whole dataset
- testset: the 35% left.

Before presenting my customized models, I desire to talk about NN and ARIMAX, on the contrary RF does not need to be newly presented since it was done in the previous section.

3.4.1 Artificial neural networks

The Artificial Neural Network was born from the idea of simulating the human nervous system's behaviour because of its undoubtedly efficient system: high learning and elaboration capacity, strong parallelization and tolerance to erroneous information. To better understand NN, let's briefly describe the human nervous system. It is composed by cells (also said neurons) linked through dendrities and axons which are fibres that allow the communication among neurons. Most of all neurons communicate only through axons in which, the communication comes as electrical impulses. When those impulses reach the presinaptic neurons, situated at the end of the axon, the latter release a chemical substance that through denditries reach the neuron(receiver). When a neuron receives a signal (information), it evaluates with an activation function whether transmits its information. To what concern artificial networks, they are commonly thought as

black boxes:

Figure 3.16: Black box idea inspired

The box has the purpose to understand and reproduce the relationships among explanatory variables (x_1,...,x_n) and response variable (y). To explain the internal mechanism of the black box, let's compare it with the biological neural network (figure 3.17). The n neurons situated as input layer represent our explanatory variables (system's features) while the weights (w_1,...,w_n) indicate in numerical terms how important is the information to deliver. Information and weigths are then linearly combined, whether its result is greater than a threshold, a function is activated. This structure is then integrated with a sub-learning algorithm which is thought to update the weigths (w_1,...,w_n) of the network. The most famous algorithm is the backpropagation, its name derives from the fact that the update comes backwards, to put it more simply once the network predicts an output, the latter is compared with the real observation and then the difference is worked out (figure 3.18 on the next page). Based on this error metric, starting from the end of the network each weight is update. This way the algorithm is adapting the model according to the experienced data, and the error metric is expected to decrease going on training. The error metric δ is the difference between the predicted value y and the observed value z, it is passed back to all the neurons contained in the previous level in order to permit them to update their weights.

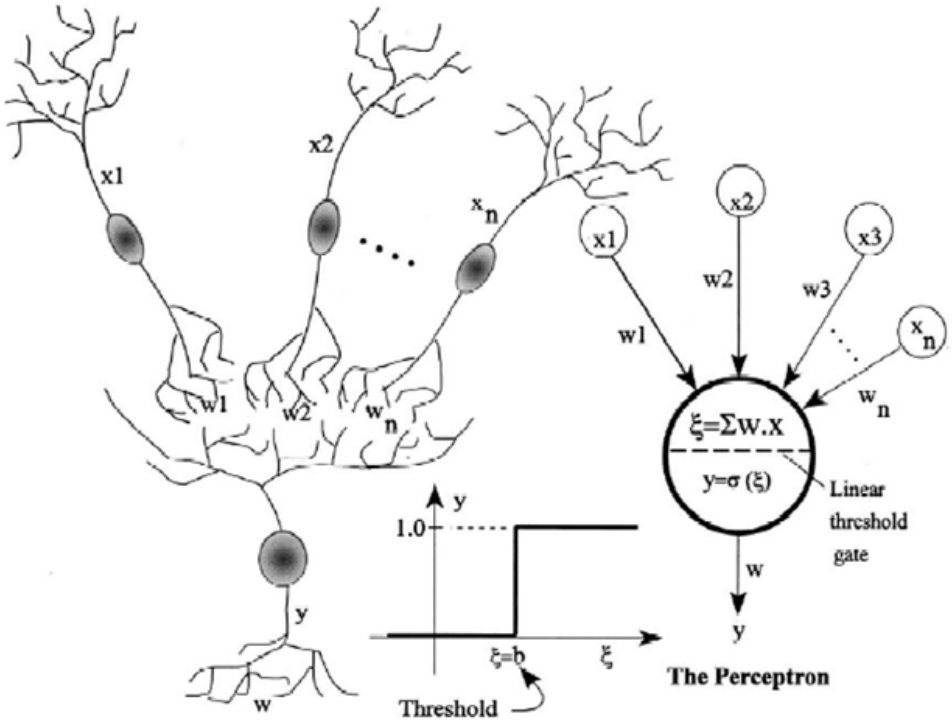


Figure 3.17: Comparison between biological and artificial neural networks, source from `http://bias.csr.unibo.it/golfarelli/DataMiningSTI/MaterialeDidattico/Reti%20Neurali.pdf`

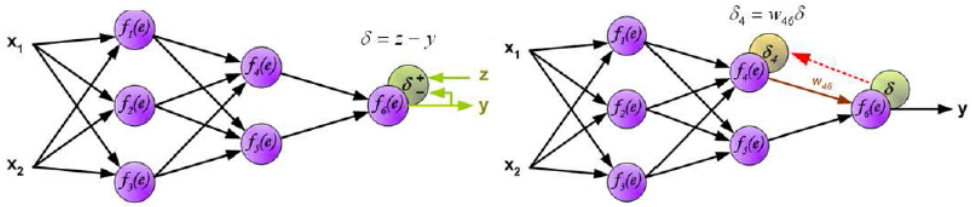


Figure 3.18: Backpropagation algorithm, source from `http://bias.csr.unibo.it/golfarelli/DataMiningSTI/MaterialeDidattico/Reti%20Neurali.pdf`

So far, I have shown only single layer NN, but there exists also Multi-layer perceptrons (MPL), they are characterized by one or more levels (hiddens) between input and output layers. This new implementation (already introduced in 3.18) is usually applied when you deal with complex tasks. MLP actually does not differ very much from single-layer NN, but it needs a further parameter to set the number of neurons inside each hidden layer. Neural network can be further categorize in:

- feed-forward network: where all the links are directed in only one direction (towards the output neuron)
- recurrent network: presence of direct cycles between two neurons

A tedious task in NNs is to set parameters so as to obtain good models, accompanied by the selection of the kind of neural network, let's list the most common parameters that characterize an artificial network:

1. feed-forward or recurrent
2. single or multi layers network (m)
3. number of nodes for each layer (h_1,...,h_n)
4. number of iteration to train the model (n)

According to the Universal approximator theorem [36] by Cybenko, I chose to implement for my three models feed-forward neural networks using only one hidden layer, I will show graphically only the neural networks related to the total electric power:

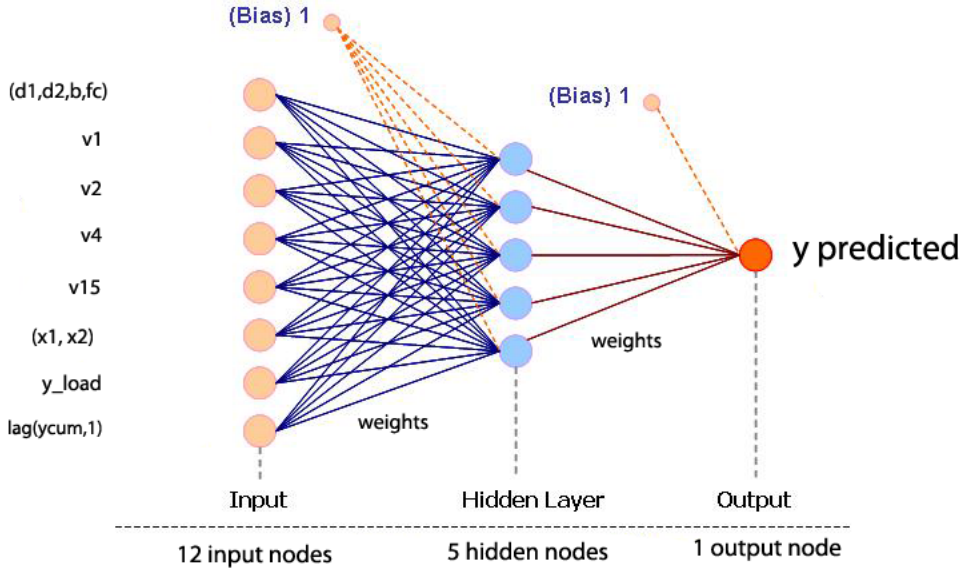


Figure 3.19: Feed-forward single layer neural network of the total electric power

Apart from the number of hidden units, graphically the three models are very similar, as a result I will not show all the three models, but I will describe their parameters according to the following notation:

$$\text{model\_params} = (\text{Input neurons, Hidden neurons , Output neurons})$$

the total electric power is then characterized by

$$y = (12,\ 5\ ,\ 1)$$

while to what concern respectively PMV and DGI, I obtained

- $x_1 = (11,\ 15\ ,\ 1)$
- $x_2 = (10,\ 10\ ,\ 1)$

To train the three models I set n=1500 iterations, since it is a good value to get convergence. To sum up it has been possible to get those parameter values testing different networks with different values and extracting those which allow me to get the best results, results have been compared using performance metrics as MAE, Std_MAE, MAPE and Std_MAPE already introduced in chapter two.

3.4.2 Time Series

In hundreads of thousands applications time series are used, it wuold be sufficient to pose more attention when we are reading a newspaper (especially an economic newspaper) time series are always present. Time Series are defined as an ordered sequence of data measured over the time, usually measured in uniform intervals. Their use ranges from economics predictions to weather forecastings. Several models have been created around time series, the most known are certainly Autoregressive processes (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) processes and ARIMA with exogenous (ARIMAX). A first class of stochastic processes is composed by AR(p), MA(q) and ARMA(p,q), these are grouped in the same set since they are dynamic linear models capable of generating stationary processes. The property of stationarity characterizes those time series with no trend (means constant mean) and constant variance. Anyway, there exists some techniques that allow to cope with non-stationarity of a series:

- differentiation filter: given a non-stationary time series W, it is posssible to get a stationary time series from it, applying the following difference

$$T_i = W_i - W_{i-d} \tag{3.10}$$

  where W_i represents the i-th data point in the W series, and W_{i-d} represents the data point placed d steps before. The parameter d can assume value one or superior, usually d is no more than two.

- Square root or logarithm can be applied only when we have stationarity in mean (with unconstant variance).

ARMA model was introduced by Box and Jenkins in 1976 at [37], some years later it has been extent in ARIMA(p,d,q) to face all those series that present non stationarity. ARIMA adds a further parameter d in its notation with respect to ARMA, d claims for differentiation technique. So far, I have been showing process generators which can model the output variable only through their past values and this is not what I desire, since I want to explain energy consumption, PMV and DGI over the selected variables. Last statement is satisfied with ARMAX(p,q,r) and ARIMAX(p,d,q,r) models, they work as ARMA and ARIMA, yet integrate the support of exogenous variables during the builging phase. In other words, the models do not depend exclusively on past values of the response variable, but even by some explanatory variables. ARIMAX is the stationary process that I selected to model my three response variables, choosing an ARIMAX rather than an ARIMA was due to the fact that I want some external variables to impact on the future predictions, for instance eventually my four controllable variables have to

modify the yield of my system. Up to now, I have generally described how to choose a model rather than another one, but I have not analyzed yet the meaning of orders p and q, not to mention the mathematical process behind each model.
The Autoregressive process of order p AR(p) identifies all those process that can be described as linear combination of p previous terms and a noyse component:

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + ... + \phi_p W_{t-p} + \varepsilon_t \tag{3.11}$$

where W_{t-p},...,W_{t-1} are the previous p term of W, while ϕ_p,...,ϕ_1 represent their respectively coefficents and ε_t is the disturbance.
Another very common stochastic process is the Moving Average of order q MA(q), which is described by a linear combination of the white noyse at the current time and in the previous q times:

$$W_t = \varepsilon_t + \sigma_1 \varepsilon_{t-1} + ... + \sigma_q \varepsilon_{t-q} \tag{3.12}$$

to what concern ARMA(p,q), it is thought as the composition of AR(p) and MA(q):

$$W_t = \sum_{i=1}^{p} \phi_i W_{t-i} + \varepsilon_t + \sum_{i=0}^{q} \sigma_i \varepsilon_{t-i} \tag{3.13}$$

Usually p and q are difficult to be individuated looking at the time series at first sight, however they are guessed using the following techniques:

1. ACF and PACF correlograms

2. feed several models with diverse parameter values, compare their results using performance metrics. Finally parameter values of the best model are adopted.

In the following lines I will talk about the first option since it was the one that I used to choose the orders of my ARIMAX according to the Box-Jenkins approach. The Correlograms are diagrams capable of showing presence of autocorelation or partial autocorrelation (respectively identified as ACF or PACF). The autocorelation function ACF indicates us whether the time series observation at time t is affected by the k-th previous observation, then expressing a memory effect, while the partial one take into account all the k previous observations.
Has shown by correlograms 3.20, the total electric power presents an evident trend since it is a cumulative variable, then it needs to be detrended. In order to obtain a detrended series I used the differencing technique, as already said one or two difference are sufficient, let's show the correlograms with only one difference in 3.21.

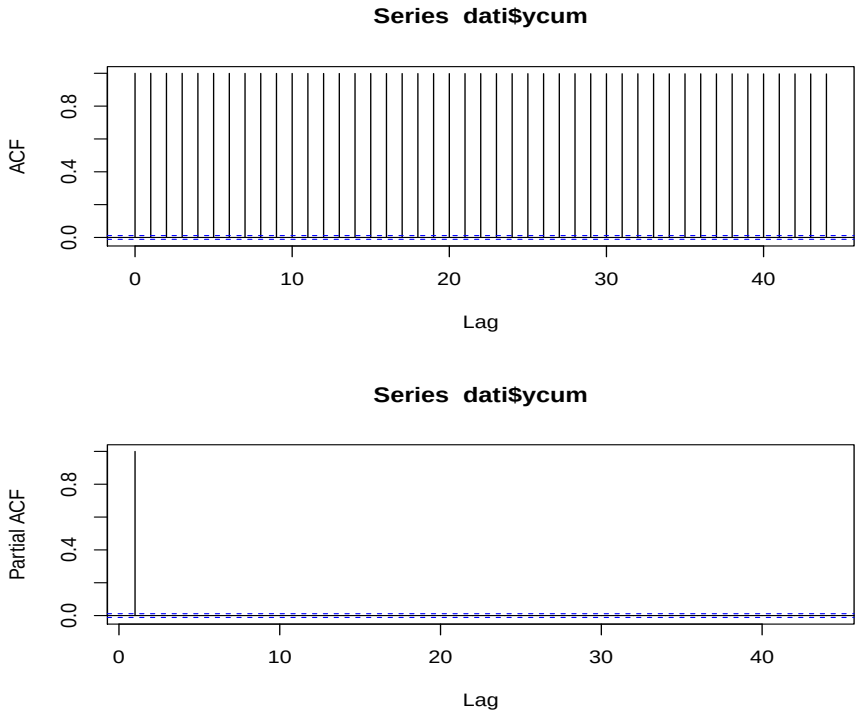


Figure 3.20: ACF and PACF correlograms related to total electric power (y)

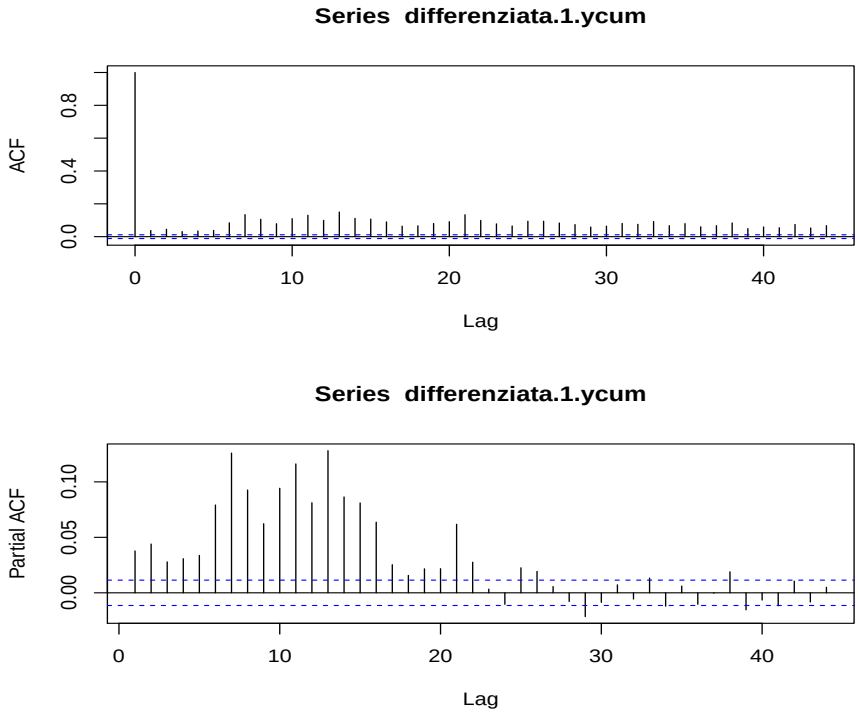


Figure 3.21: ACF and PACF correlograms related to total electric power (y) differenced by one (d=1)

To obtain a stationary time series we expect that one of the two correlograms present a tail which gradually get zero. In figure 3.21 no descending tails are present neither in ACF nor in PACF, therefore another difference can be applied. As expected in figure 3.22 PACF correlogram displays autocorrelations with previous lags gradually get zero, then we are surely dealing with a stochastic process that present an MA component, to spot its order is sufficient to look at the ACF correlogram: the number of lags which are not contained in the confidence interval represents the order q. Doubtless q = 1, do not let you get confused by the autocorrelation at time zero which must not be considered.
To sum up the total electric power has been modelled with an ARIMAX(p = 0, d = 2, q = 1) model.

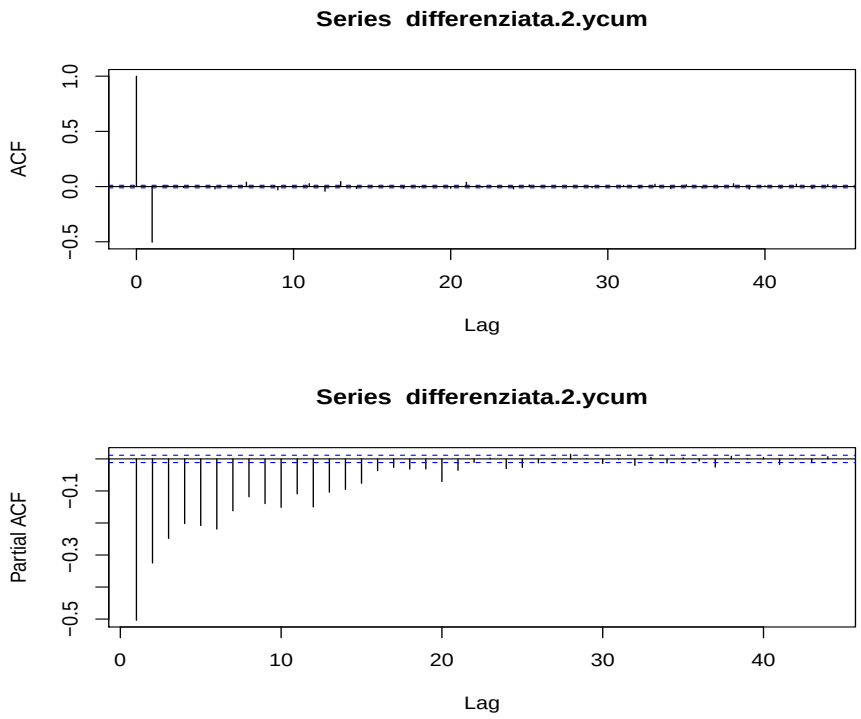


Figure 3.22: ACF and PACF correlograms related to total electric power differenced by two (d=2)

To what concern the thermal comfort, PMV needs to get stationary, thus differencing at most twice it should be possible to obtain constant mean and constant variance.

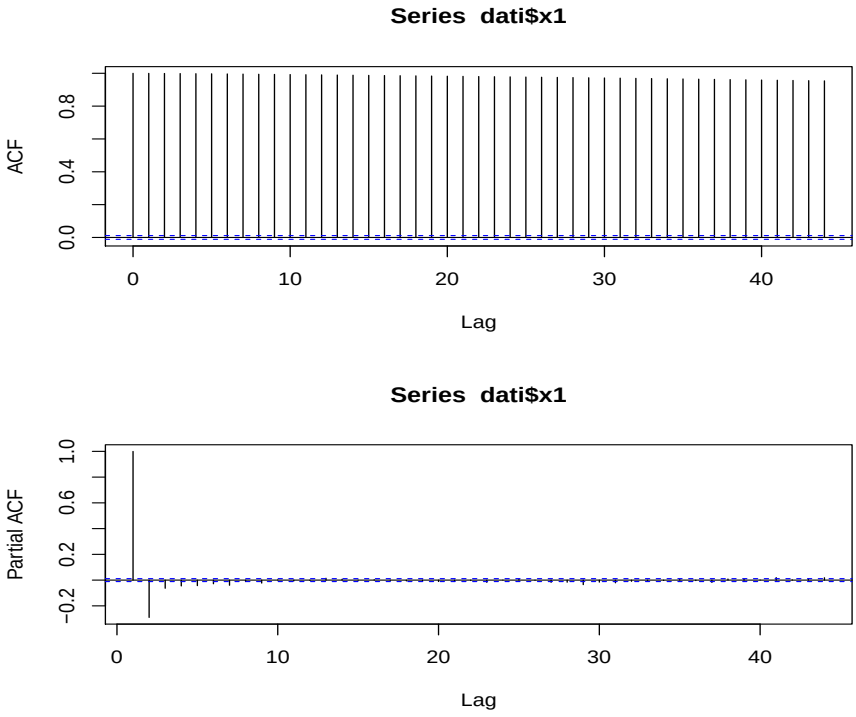


Figure 3.23: ACF and PACF correlograms related to PMV

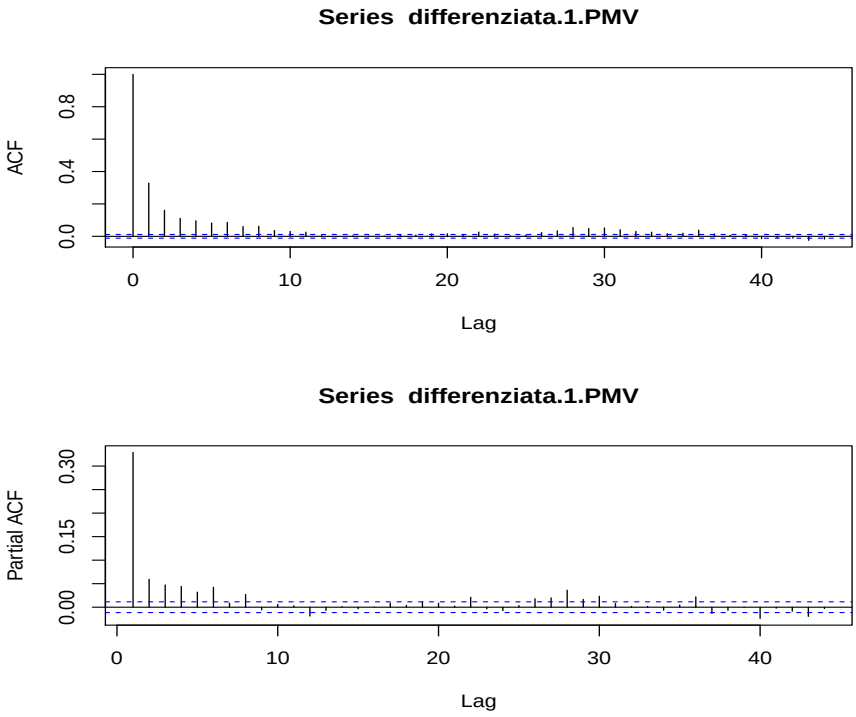


Figure 3.24: ACF and PACF correlograms related to PMV differenced by one (d=1)

In figure 3.25, it is confirmed that stationarity can be obtained differencing PMV twice. Likewise it happens in total electric power an MA stochastic process should be used, the figure suggests me to use a second order for q, however a first order wuold be sufficient since the second lag in the autocorrelation diagram is inferior to -0.1. Hence I chose to use an ARIMAX(p = 0, d = 2, q = 1).

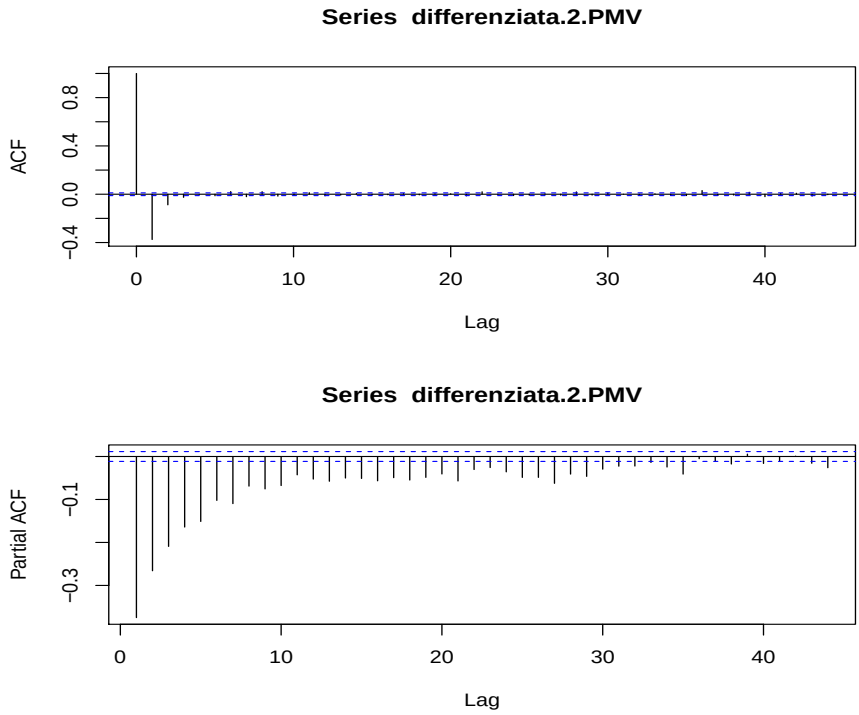


Figure 3.25: ACF and PACF correlograms related to PMV differenced by two (d=2)

On the contrary of total electric power and PMV, in figure 3.26 it is evident as DGI no needs of differencing, since a tail that gradually get zero is present in the autocorrelation correlogram, thus this time an autoregressive component of order 2 is present as it is possible to evince from the PACF correlogram. Hence, an ARIMAX(p = 2, d = 0, q = 0) model is necessary.

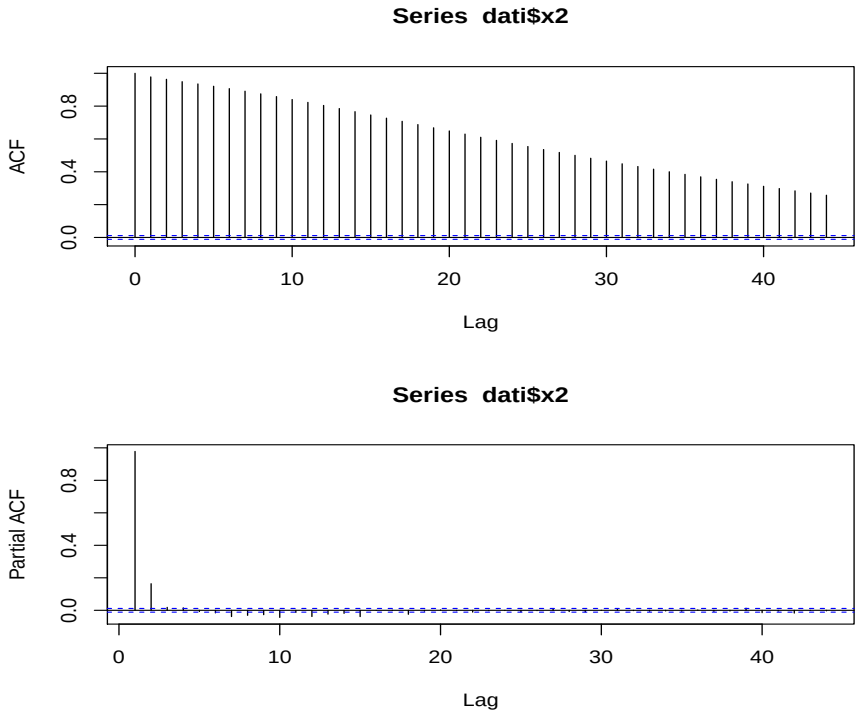


Figure 3.26: ACF and PACF correlograms related to DGI

3.4.3 Random Forest

Random Forest is a combination of weaker trees whose predictions are finally aggregated through an average prediction value in regression problem (see figure 3.4.3), whereas a majority vote is adopted in classification problems.

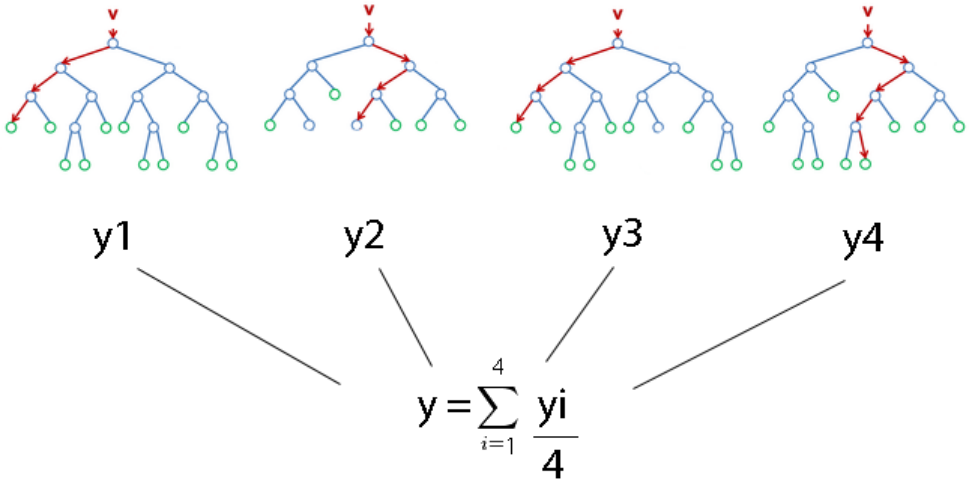


Figure 3.27: Example of Random Forest for regression problems

This section has been added not to newly describe RF since it has been already done in

section 3.3, but to recall you that RF was also used to model energy consumption, PMV and DGI. For practical use RF requires an accurate settings of the following parameters: the number of binary trees to generate (n), the number of variables randomly selected as candidates at each splitting phase (m) and the depth of the tree (d). According to [39], we have set the following RF parameters (n = 1000, m = p/3, d = maximal) where p is the number of explanatory variables selected by feature selection process. Regards the depth of each tree we chose to allows RF to generate maximal tree size.

3.5 Model Validation

Why is the evaluation of each model considered so important after the modelling phase? Mainly a couple of points can be listed:

- to determine the accuracy degree of a model with respect to the real system
- to select the best model among several ones, the one which better fits our data

The first point highlights the fact that very often modelling is applied to test particular conditions that cannot be tested in reality, for instance shooting a rocket on the space is a complex task, it becomes clear as testing the accuracy of the model which simulate the rocket's behaviour is necessary. Using an inaccurate model to simulate it, could mean a catastrophic failure in reality. The second point enhances our job, since permits us to select the model which better simulate the rocket's behaviuor.
American Society of Mechanical Engineers (ASME) defined model validation as "process of determining the degree to which a computer model is an accurate representation of the real world from the perspective of the intended model applications." A motivational work on model validation is performed by Paez [38]. To calculate the accuracy of a model, the decision maker must select some performance metrics which will compare the model predictions with the measured values, finally providing a mechanism to confront several models. I decided to use those metrics used by Kusiak in order to maintain a connection with its outstanding work: MAE, Std_AE, MAPE and Std_APE. Having already introduced them in chapter 2, I can move on directly to the results obtained by NN, ARIMAX and RF during 6 hours predictions performed in November 27th, 2013:

Model	MAE	Std_AE	MAPE	Std_APE
NNet	0.0133170931	0.0022563158	0.0001086740	0.0000183701
ARIMAX	0.731091091	0.343465180	0.005958933	0.002792062
RF	1.873719682	0.618466976	0.015279428	0.005020644

Table 3.1: Performance metrics of the three total electric power models on 6 hours predictions

Model	MAE	Std_AE	MAPE	Std_APE
NNet	0.0003140707	0.0002924790	0.0003796704	0.0002908663
ARIMAX	0.012038623	0.009457512	0.015536649	0.007960794
RF	0.007979752	0.006511992	0.012287685	0.009703844

Table 3.2: Performance metrics of the three PMV models on 6 hours predictions

Model	MAE	Std_AE	MAPE	Std_APE
NNet	0.0001638867	0.0001183658	0.0002538736	0.0006446201
ARIMAX	3.1980896	0.8735897	4.5334896	3.7888799
RF	0.01922089	0.04763736	0.08496009	0.58551323

Table 3.3: Performance metrics of the three PMV models on 6 hours predictions

The performance metrics have been worked out on 72 predictions (equivalent to 6 hours), it results that neural network is certainly the model that better fits to model all three different components: total electric power, PMV and DGI. The next plots will graphically confirm my last statement (see figures 3.28 on the following page, 3.29 on the next page and 3.30 on page 83). Each plot presents in the x and y axis respectively the daily hours (related to November 27th, 2013) and the response variable, the presence of a dash line in the graphs marks the moment from which is performed the prediction (right hand side) from the preceding observations (left hand side). These figures are definitely interesting since they allow me to see the previous behaviour of the system and the predictions of my models with respect to the real system's measurements in the 6 hours predictions. Observing the figures is now more evident as the NNet prediction lines overlap the real measurements line, approving that NNet seems to be the most qualified to model our response variables. Anyway good results are also obtained by RF and ARIMAX to model PMV, in case they could be considered as second choices (see figure 3.29). In contrast to what happen with energy and PMV, to note that ARIMAX gets the worst results in predicting the daylight glare index (figure 3.30), while quite good predictions are achieved by RF.

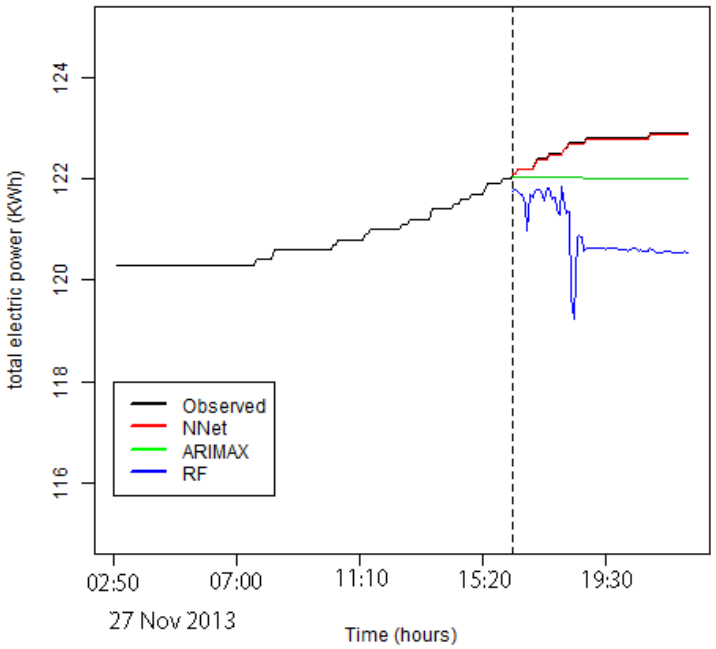


Figure 3.28: Energy predictions related to NN, ARIMAX and RF in 27th November 2013

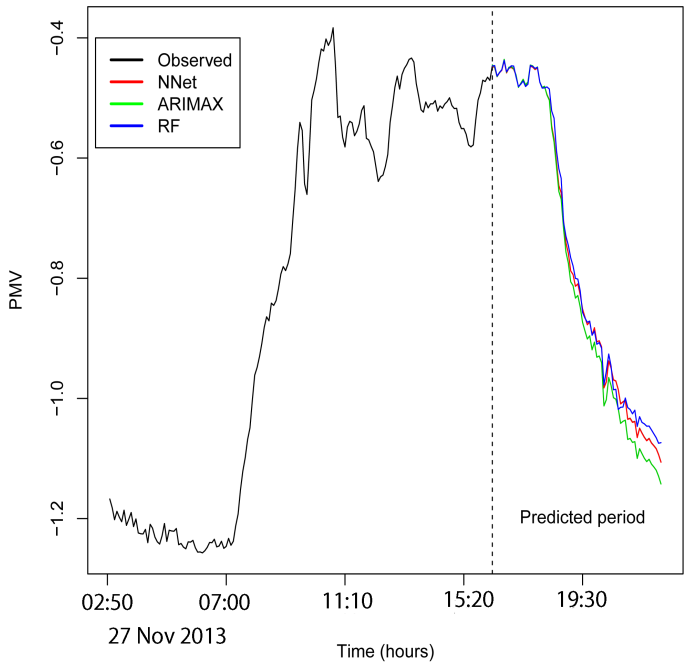


Figure 3.29: PMV predictions related to NN, ARIMAX and RF in 27th November 2013

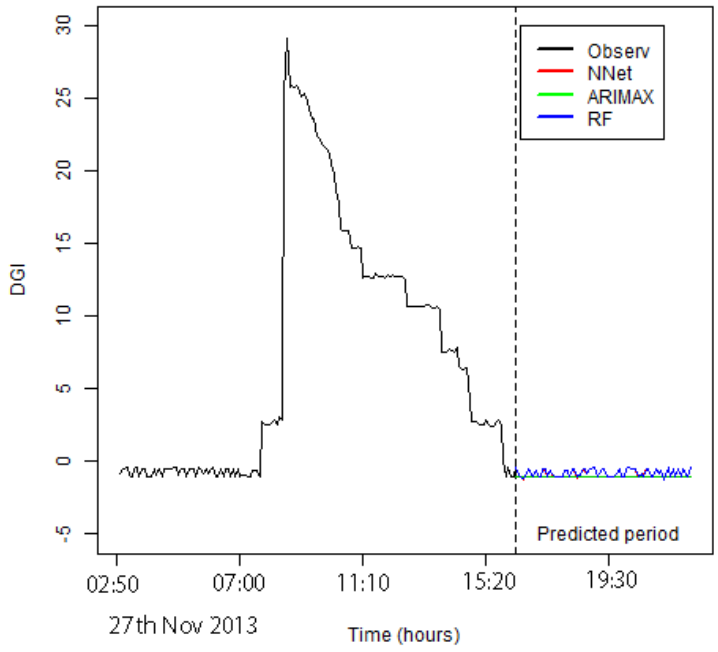


Figure 3.30: DGI predictions related to NN, ARIMAX and RF in 27th November 2013

3.6 Optimization

Optimization is a fundamental procedure in many business and engineering applications,however in everyday life our activities are often scheduled according to an optimization process, for example when we select the vehicle to go to the hairdresser's sometimes we need to save our time so our choice will be either a car or a motorcycle, other times the criteria will be to save energy (fuel in this case) then we will opt for the evergreen bike. Anyway single objective optimization problems are seldom present in engineering problems, in fact they often require for multiple objectives, which are sometimes conflicting among them. To clear the term conflicting, let's introduce another example: the problem consists of determining the best transport vehicle based on two criterias:

1. affordable distance in a day
2. energy consumption

Of course our two criterias of choice represent our objective functions, the fact that they are conflicting means they are in contrast of interests since maximizing the affordable distance results in consuming more energy and viceversa. As a result finding a unique solution (the best one) straightforward is not a trivial task. Consequently a set of acceptable optimal solutions is required in multi-objective optimization, on the contrary in a single-objective problem a unique solution is looked for. This new set can be considered as a trade-off set and can be turned into an advantage for the decision makers (DM) since it allows them to choose, among a set, the most appropriate

solution for their problems. From now, this set will be called Pareto set.
Considering the transport problem, reasonably we could rank all our feasible solutions according to the two criterias, follows the figure:

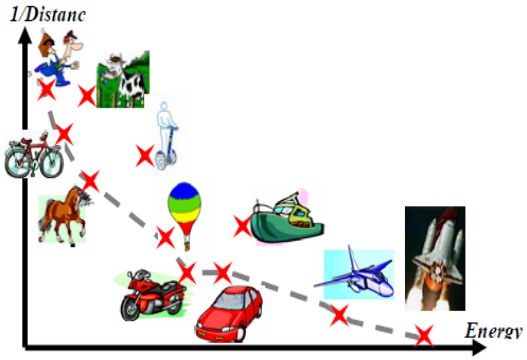


Figure 3.31: Transportation mode example

Those vehicles placed on the Pareto curve represents the Pareto set, the DM will choose the most appropriate vehicle according to its necessity.
In order to explain the Pareto approach, let's formulate mathematically a multi-objective problem:

$$\begin{aligned} Min \quad & F(\overrightarrow{x}) = [f(\overrightarrow{x_1}), ..., f(\overrightarrow{x_n})] \\ & subject \quad to \qquad h_k(\overrightarrow{x}) \leq 0 \quad k = 1, .., m \\ & where \quad \overrightarrow{x} = (x_1, ..., x_n) \end{aligned} \tag{3.14}$$

where $f(\overrightarrow{x_1}), ..., f(\overrightarrow{x_n})$ represent our conflicting objective functions, while $\overrightarrow{x}$ identifies a combination of our n decision variables, and $h_k(\overrightarrow{x})$ are constraints.
The Pareto method will find a set of non-dominated $\overrightarrow{x}$ solutions, to put it more simply it means no other solutions can improve an objective without degrading at least another one. Non-dominated solution are a subset of all the feasible one, with feasible it is intended all those solutions that satisfies the constraints $h_k(\overrightarrow{x}) \leq 0$.
Having introduced the term non-dominated solution, it is now time to formalize the notion of Pareto-dominance: in a bi-objective minimization problem, given two feasible solutions $\overrightarrow{a} = (\overrightarrow{a_1}, ..., \overrightarrow{a_n})$ and $\overrightarrow{b} = (\overrightarrow{b_1}, ..., \overrightarrow{b_n})$, $\overrightarrow{a}$ dominates $\overrightarrow{b}$ if and only if

$$f_1(\overrightarrow{a}) \leq f_1(\overrightarrow{b}) \quad and \quad f_2(\overrightarrow{a}) < f_2(\overrightarrow{b}) \tag{3.15}$$

equation [**?**] can be generalized in a multi-objective context as:

$$\forall i \quad \in 1, .., n \quad f_i(\overrightarrow{a}) \leq f_i(\overrightarrow{b}) \quad and \quad f_j(\overrightarrow{a}) < f_j(\overrightarrow{b}) \tag{3.16}$$

Hence, a feasible solution belongs to the Pareto set if no other solutions dominates it. Graphically a Pareto set is often depicted as a surface of the search space that contains the non-dominated optimal solutions of the problem. To better conceive the last concept, consider two non-conflicting objective functions, which needs to be minimized, follows a figure:

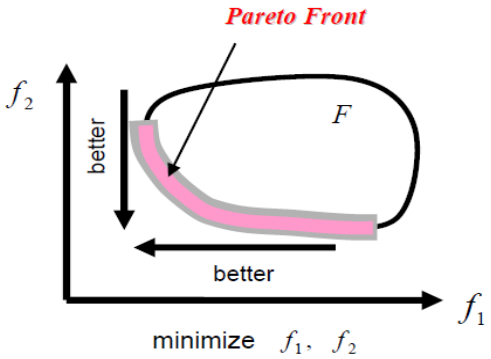


Figure 3.32: Pareto front shape

Apparently the black line identifies where the feasible solutions are placed, while the optimal solutions are those points closer to the axis intersection, graphically lied in the pink surface, that surface is called Pareto front. Different approaches can be implemented to solve MO problems, in particular I will talk about:

- classical methods
- Pareto front through intellingent techniques

Classical approach starts from the idea that a multi-objective problem is a generalization of a single-objective one, therefore it is possible to convert the first one into the second one typically aggregating the multi objectives into an unique function to optimize. Additionally, this method removes the idea to find a set of optimal solutions, limiting the search to find straightforward a solution. Usually a linear weighted combination of n functions is applied:

$$min \qquad F = \sum_{i=0}^{n} w_i f_i(x_i) \quad with \quad w_i \geq 0 \quad and \quad \sum_{i=0}^{n} w_i = 1 \tag{3.17}$$

This technique is suggested when a priori information is owned by the decision maker, because of the weights (w_1,...,w_n) that must be chosen appropriately otherwise our optimization could be erroneously affected. A linear weighted combination was used also by Kusiak at [27], [28], [29] even if he used it for different purposes, in fact he grouped all constraints into one, weighting each of them so as to pay more attention to some constraints rather than others in different period of the year. For example, in winter more attention should be given to room temperature constraint rather than room humidity because of the low humidity presence, common in winter season.
The second approach is pushed by the idea of finding the Pareto set. Today thanks to the increment of power computation this technique is often combined with metaheuristic algorithms, for instance Kusiak implemented SPEA (Strenght Pareto Evolutionary Algorithm) to find the Pareto front using an evolutionary algorithm at [28]. Other intellingent techniques to find the Pareto set have been described at chapter 1 such as Genetic algorithms and Particle Swarm, so in this section they do not need further explanations. To be coherent with the scientific work, I formulated mathematically the Stra.t.e.g.a problem as a constrained multi-objective problem:

$$\min_{(d_1,d_2,b,f_c)} y$$
subject to
$$y = f(d1, d2, b, fc, v1, v2, v4, v15, y\_load, x1, x2, lag(y,1))$$
$$x1 = f(d1, d2, b, fc, v1, v2, v4, v7, v11, y\_load, lag(x1,1))$$
$$x2 = f(d1, d2, b, fc, v2, v5, v6, v11, y\_load, lag(x2,1))$$
$$-0.5 \leq x_1 \leq 0.5$$
$$x_2 \leq 21$$

where the lagged variables represent the variable values at previous time, in our case 5 minutes before. The formulation describes perfectly our final objective that I recall you in order to avoid misunderstanding. The purpose is to find the best future configuration settings of (d_1, d_2, b, f_c) that allow me to minimize the energy consumption (y), and at the same time maintaining good levels of comfort as PMV and DGI (x_1, x_2) expressed mathematically as constraints.
In the next figure I will show the optimization process algorithmically speaking:

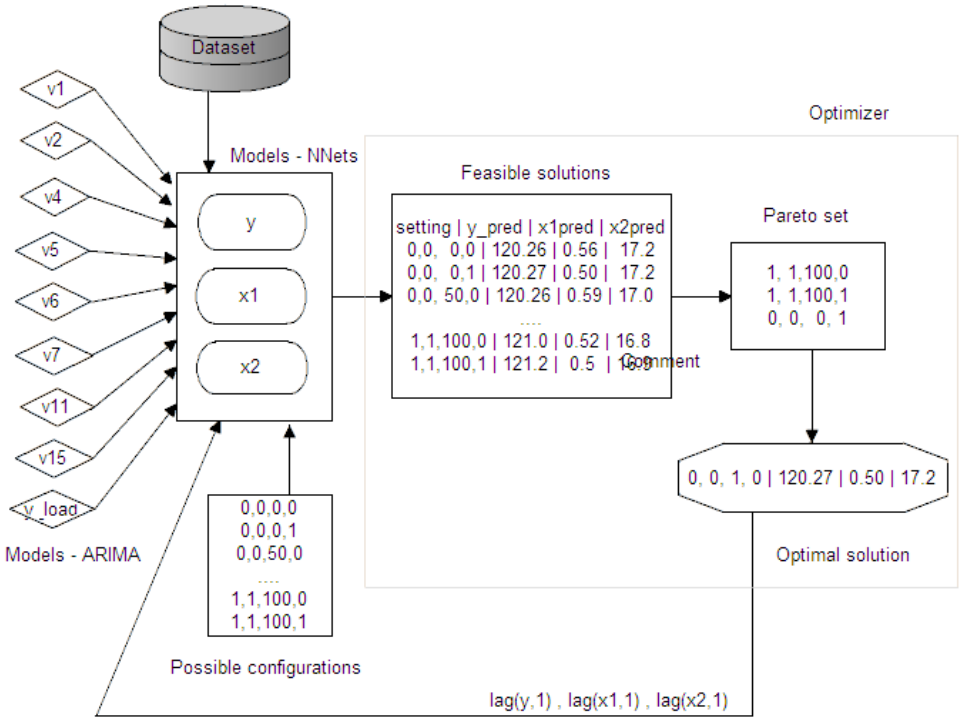


Figure 3.33: Optimization process

My optimizer is an Exhaustive Pareto Optimization algorithm (EPO) since it exploits an exhaustive search, capable of extracting the Pareto set and choose the best optimal solution. Let's summarize step-by-step how it works:

1. The uncontrollable variables (v1, v2, v4, v5, v6, v7, v11, v15, y_load) are modeled exploiting ARIMA models, in fact we do not expect them to depend on other variables, but to depend on their past, therefore time series modeling is surely the most adapt to understand the uncontrollable variables' behaviour over the time. In order to avoid to be repetitive, in this

section we will not demonstrate how to spot p, q and d orders of each ARIMA model since the approach to do it has been already shown in section 3.4.2.

2. The three neural networks (y, x1, x2) are then fed by ARIMA forecastings, previous lags and alternatively by each possible configuration of the controllable variables, therefore obtaining a prediction for energy, PMV and DGI for each possible setting implemented. Here, an exhaustive search has been reasonable since the possible configurations are only 24 (given by the product of the possible values assumed by d_1, d_2, b, f_c, then $2 \cdot 2 \cdot 3 \cdot 2 = 24$). Such exhaustive search wuold be unpracticable to use if I had hundreads of thousands combinations, in that case it wuold be suitable an intelligent technique such as GA.

3. From the set of 24 predictions (feasible solutions) is then extracted a Pareto set, to get those solutions which are non-dominated by others.

4. Obtained the Pareto set, the optimizer needs to choose the best solution: the one which minimizes the energy consumption.

5. From the optimal solution are then extracted the prediction of y, x1 and x2 and passed to the next iteration as previous lags of total electric power, PMV and DGI. Skip to instruction 1.

To what concern the selection of solutions which belong to the Pareto set at each iteration, follow a simple algorithm :

1. an Archive is created to contain all potential predictions (24 possible settings)

2. the Archive is then order according to DGI (which needs to be minimized) in an ascending way

3. the first solution is added to the Pareto set

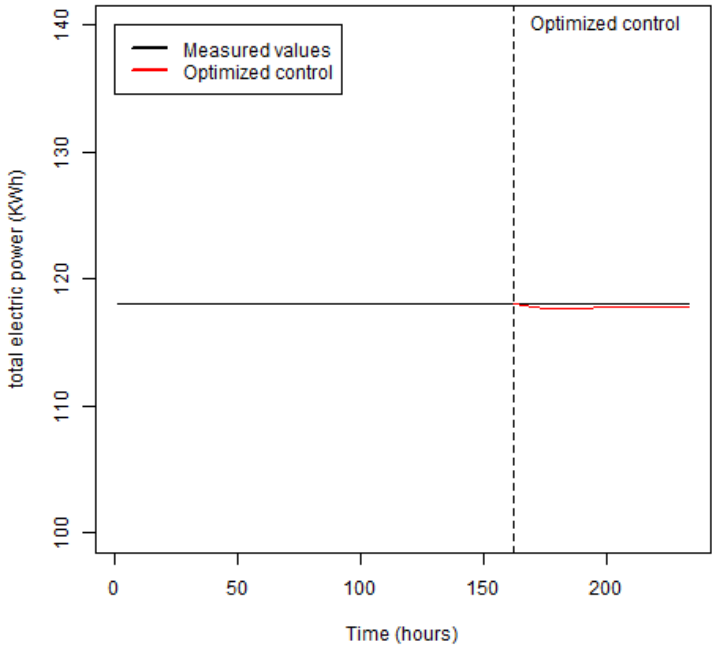
4. cycle (to scan all the 24 solutions)

    a) if the current solution has PMV inferior to all the Pareto-dominant solutions then add the current solution to the Pareto set, else go on.

    end_ciclo

The proposed Exhaustive Pareto Optimization algorithm is thought to work out the optimization model every five minutes. As a result we could hypothetically expect also to see different control settings during all the time prediction. To show the extraordinary potential of my optimizer, let's suppose to perform 6 hours prediction starting from a randomly chosen time point of the test set. I want to compare the observed values of total electric power, PMV and DGI with respect to their respectively optimized predictions. The optimized predictions represent the hypothetical behaviour of the system adjusted by the control settings selected by EPO. Figures 3.34, 3.35, 3.36 are characterized by a vertical dash line (at point 161) which separates the system behaviour till that moment from the prediction period. Regards PMV and DGI the figures 3.35, 3.36 will

be further horizontally marked by the limits of wellness condition ($-0.5 \leq PMV \leq 0.5$ and $DGI \leq 21$).



5.

Figure 3.34: The observed and optimized total electric power prediction

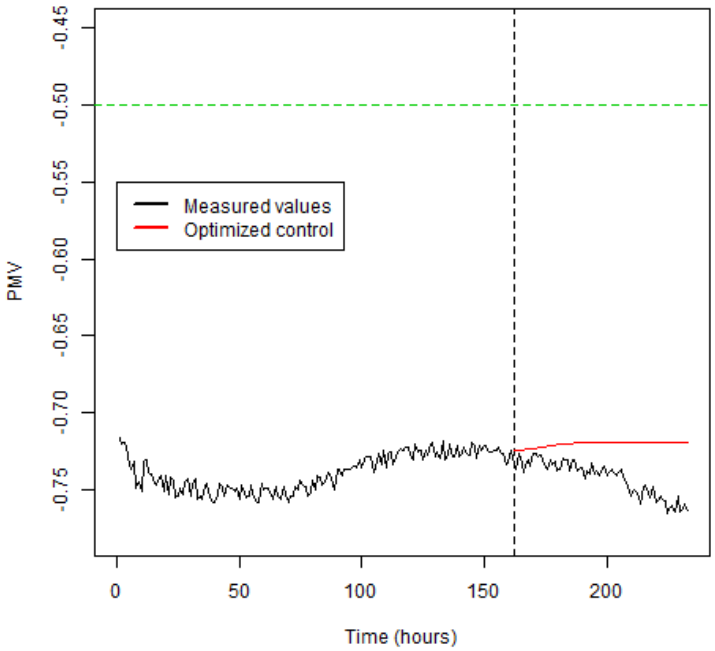


Figure 3.35: The observed and optimized PMV prediction

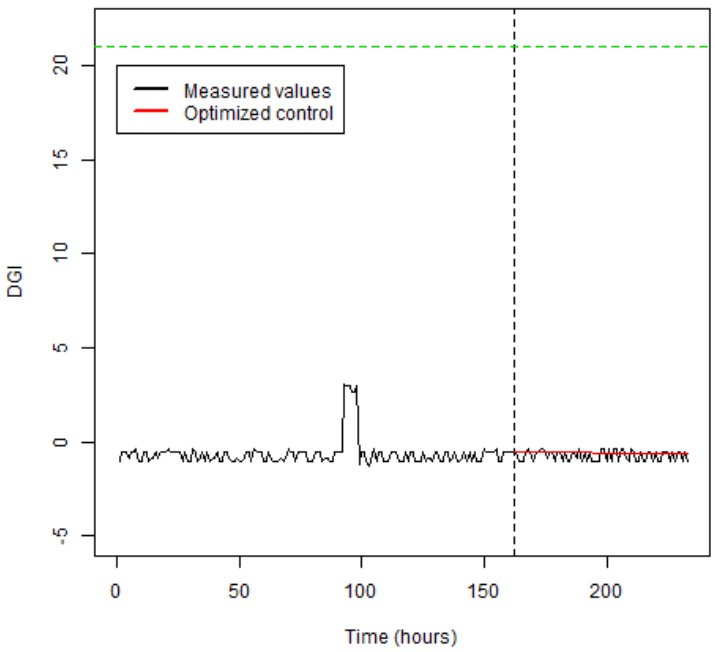


Figure 3.36: The observed and optimized DGI prediction

Figure 3.34 shows that measured values (basically the observed behaviour of the system) are constant during the all considered period of 233 observations, therefore the system is definetely not consuming maintaining the same configuration settings as shown in table 3.4 (left table). However, the observed PMV values are violating the constraint on PMV (see figure 3.35), furthermore its trend is still slowly decreasing at the first prediction point, so EPO has the objective of increasing the thermal comfort, moving the PMV towards the best wellness condition (close to zero). In tables 3.4 are shown respectively the control settings selected by the system and those chosen by the optimizer. To highlights the fact that the system is prone to maintain the same configuration for the prediction period, whereas in the first 15 intervals the optimizer change a couple of configurations, after that it gets the system stable through an unique configuration.
Now looking at the figure 3.34 it is possible to see the outstanding results obtained in the six hours prediction by EPO, in fact the optimized control allows the system not to consume more energy than the real system, what is more the optimized PMV tends to get near the wellness condition. To note that the optimized total electric power tends to decrease during the prediction period, which is not possible in reality since it is a cumulative variable, anyway in this case it is due to a prediction error. To what concern DGI (figure 3.36) the optimized control maintains the system in a wellness condition, so it does not need to either increase or decrease.

obs	d1	d2	b	fc
1	0	0	100	0
2	0	0	100	0
3	0	0	100	0
4	0	0	100	0
5	0	0	100	0
6	0	0	100	0
7	0	0	100	0
8	0	0	100	0
9	0	0	100	0
10	0	0	100	0
11	0	0	100	0
12	0	0	100	0
13	0	0	100	0
14	0	0	100	0
15	0	0	100	0
16	0	0	100	0
...	0	0	100	0
...	0	0	100	0
71	0	0	100	0
72	0	0	100	0

obs	d1	d2	b	fc
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0
5	0	1	0	0
6	0	1	0	0
7	0	1	0	0
8	0	1	0	0
9	0	1	0	0
10	0	1	0	0
11	0	1	0	0
12	0	1	0	0
13	0	0	0	1
14	0	1	0	0
15	0	0	0	1
16	0	0	0	1
...	0	0	0	1
...	0	0	0	1
71	0	0	0	1
72	0	0	0	1

Table 3.4: Control settings selected respectively by the system and the optimizer

Chapter 4

Conclusions and future developments

A data driven approach was employed to optimize a real HVAC system. The real case study is a building composed by three offices, in particular I applied my solution to office 2 which was designed to be completely automatic. The objective of this thesis is to find the optimal configuration of controllable variables that maximize energy savings taking care of comfort levels for inhabitants such as thermal comfort (PMV index) and daylight glare (DGI).
Random Forest and a correlation analysis were initially applied to select the most significant features to construct three prediction models for total electric power, PMV and DGI. To build the models were trained a neural network, an ARIMAX model and a Random Forest. After a validation process a feed forward single-layer neural network was chosen since it outperforms ARIMAX and Random Forest in terms of prediction for all my three objectives.
A multiobjective optimization model was then formulated and solved by an Exhaustive Pareto Optimization algorithm. The Exhaustive Pareto was set to search for the optimal control solution every five minutes so as to minimize the total electric power preserving tolerable levels of comfort.
I wuold like now to talk about prons and cons of my proposal. Concerning the structure of my implementation the most evident advantage is that it was designed to satisfy the modularity property, in fact it can be decomposed mainly in three modules: feature selection, modeling and optimization. This structure will allow future developers to implement their customized modules, for instance instead of using neural networks in modeling phase, a Support Vector Machine could be employed. Another important strenght is due to the fact that those solutions are not still implemented at all in buildings, let's imagine to implement in public buildings such as hospitals and schools a similar automatic tool, certainly we will reduce gas emissions in air, greenhouse effect, acid rains in terms of environmental issues, while in terms of money we will minimize our public outcomes.
The principal advantage of my solution is the computing time, currently a six hours prediction takes a couple of minutes to be performed, then a single day estimation wuold take approximately ten minutes. With regard to future developments: at present my solution implements an Exhaustive Pareto optimization which evaluates each feasible configuration to get the best one, but this is affordable till the number of possible configurations are in the order of hundreads or thousands. Different it wuold be whether we had hundreads of thousands, in this case an intelligent algorithm such as genetic algorithm or particle swarm optimization wuold be preferred

in order to get near optimal solutions. Obviously a parallelized solution should be integrated to obtain good speed performance. Furthermore my optimization algorithm currently exploits DGI and PMV to get the Pareto set, in future further comfort metrics could be easily integrated such as indoor air quality.

Bibliography

[1] Steemers, K., *Energy and the city: density, buildings and transport*, in Energy and Buildings, 35, 2003, 3–14.

[2] International energy agency, *Worldwide Trends in Energy Use and Efficiency*, Technical report, 2008, http://www.iea.org/publications/freepublications.

[3] Perez-Lombard, L., Ortiz, J., Pout, C., *A review on buildings energy consumption information*, in Energy and Buildings, 40, 2008, 394–398.

[4] Shafiee, S., Topal, E., *When will fossil fuel reserves be diminished?*, in Energy Policy, 37, Issue 1, January 2009, 181–189.

[5] Fanger, P.O., *Thermal Comfort*, McGraw-Hill Inc., New York, USA, 1970.

[6] Gameiro da Silva, M. C., *Spreadsheet for the calculation of thermal comfort indices PMV and PPD* , Technical report, Department of mechanical engineering - university of Coimbra, Coimbra, Portugal, 2013.

[7] Tugrul Ogulata, R., *The Effect of Thermal Insulation of Clothing on Human Thermal Comfort*, FIBRES & TEXTILES in Eastern Europe, 15, No. 2 (61), 2007, Turkey.

[8] Bellia, L., Cesarano,A., Iuliano, G.F., Spada, G., *Daylight glare: a review of discomfort indexes*, Energy, 36, 2011, 5935-5943.

[9] Nazzal, A. A., *A new daylight glare evaluation method: Introduction of the monitoring protocol and calculation method*, Energy and Buildings, 33, 2001, 257-265.

[10] Settimo, G., Brini, S., Baldassarri L.T., De Martino, A., Lepore, A., Moricci, F., *Presenza di CO_2 e H_2S in ambienti indoor-residenziali: analisi critica delle conoscenze di letteratura*, Technical report, Istituto superiore di sanità, Rome, Italy, http://www.iss.it/binary/iasa/cont/CO2_H2S_FINALE.pdf.

[11] Collette, Y.,*Multiobjective optimization: principles and case studies*, Phd thesis, 2003, page 6.

[12] Gass, S.I., *Linear Programming: Methods and Applications*, 5th ed., McGraw-Hill, 2003, New York, US.

[13] Zhao, H. X., Magoulès, F.,*A review on the prediction of building energy consumption*, Renewable and Sustainable Energy Reviews, 16, 2012, 3586-3592.

[14] ISO, *Energy performance of buildings? Calculation of energy use for space heating and cooling*,2008, Geneva, Switzerland.

[15] Yao, R., Steemers, K.,*A method of formulating energy load profile for domestic buildings in the UK*, Energy and Buildings, 37, 2005, 663-671.

[16] Al-Homoud, MS.,*Computer-aided building energy analysis techniques*, Building and Environment, 36, 2001, 421-433.

[17] Wang, S., Xu , X.,*Simplified building model for transient thermal performance estimation using GA-based parameter identification*, International Journal of Thermal Science,45, 2006, 419-432.

[18] Mustafaraj, G., Chen, J., Lowry, G., *A Development of room temperature and relative humidity linear parametric models for an open office using BMS data*, Energy and Buildings, 42, 2010, 348-356.

[19] Kimbara, A., Kurosu, S., Endo, R., Kamimura, K., Matsuba, T., Yamada, A., *On-line prediction for load profile of an air conditioning system*, ASHRAE Transactions, 101, 1995, 198-207.

[20] Kimbara, A., *Peak demand control in commercial buildings with target peak adjustment based on load forecasting*, Proceedings of the 1998 IEE international conference on control applications, 2, 1998, 1292-1296.

[21] Ooka, R., Komamura, K., *Optimal design method for building energy systems using genetic algorithms*, Building and Environment, 44, 2009, 1538-1544.

[22] Kalogirou, SA., Neocleous, CC, Schizas,CN., *Building heating load estimation using artificial neural networks*, Proceedings of the 17th international conference on paralleal architecture and compilation techniques, 1997.

[23] Kreider, JF., Claridge, DE.,Curtiss, P., Dodier, R., Haberl, JS., Krarti, M.,*Building energy use prediction and system identification using recurrent neural networks*, Journal of solar energy Engineering, 117, 1995, 161-166.

[24] Magnier, L., Haghighat, F, *Multiobjective optimization of building design using TRNSYS simulations, genetic algorithm, and Artificial Neural Network*, Building and Environment, 45, 2010, 739-746.

[25] Chow, T.T., Zhang, G.Q., Lin, Z., Song, C.L., *Global optimization of absorption chiller system by genetic algorithm and neural network*, Energy and Buildings, 34, 2002, 103-109.

[26] Kusiak, A., Li, M., Tang, F., *Modeling and optimization of HVAC energy consumption*, Applied Energy, 87, 2010, 3092-3102.

[27] Kusiak, A., Xu, G., *Modeling and optimization of HVAC systems using a dynamic neural network*, Energy, 42, 2012, 241-250.

[28] Kusiak, A., Tang, F., Xu, G., *Multi-objective optimization of HVAC system with an evolutionary computation algorithm*, Energy, 36, 2011, 2440-2449.

[29] Kusiak, A., Xu, G., Tang, F., *Optimization of an HVAC system with a strenght multi-objective particle-swarm algorithm*, Energy, 36, 2011, 5935-5943.

[30] Kusiak, A., Zeng, Y., Xu, G., *Minimizing energy consumption of an air handling unit with a computational intelligence approach*, Energy and Buildings, 60, 2013, 355–363.

[31] Tukey, J. W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

[32] Breiman, L., *The Boosting Approach to Machine Learning: An Overview*, Nonlinear Estimation and Classification, Springer, 2003.

[33] Breiman, L., *Bagging predictors*, Machine Learning, 21, 1996, 123-140.

[34] Breiman, L., *Random Forests*, Machine Learning, 45, 2001, 5-32.

[35] Pradeep, K., Dadhich, A. *Study of correlation between Human height and foot length in residents of Mumbai*, International Journal of Biological & Medical Research, 3, 2012, 2232-2235.

[36] Cybenko, G., *Approximations by superpositions of sigmoidal functions*, Mathematics of Control, Signals, and Systems, 1989.

[37] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., *Time Series Analysis: Forecasting and Control*, John Wiley & Sons Inc., New York, 2008.

[38] Paez, T.L., *Introduction to Model Validation*, Validation and Uncertainty Quantification Department, Sandia National Laboratories, Albuquerque, New Mexico, 2009.

[39] Liaw, A., Wiener, M., *Classification and Regression by Random Forest*, 2002, from `http://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf`