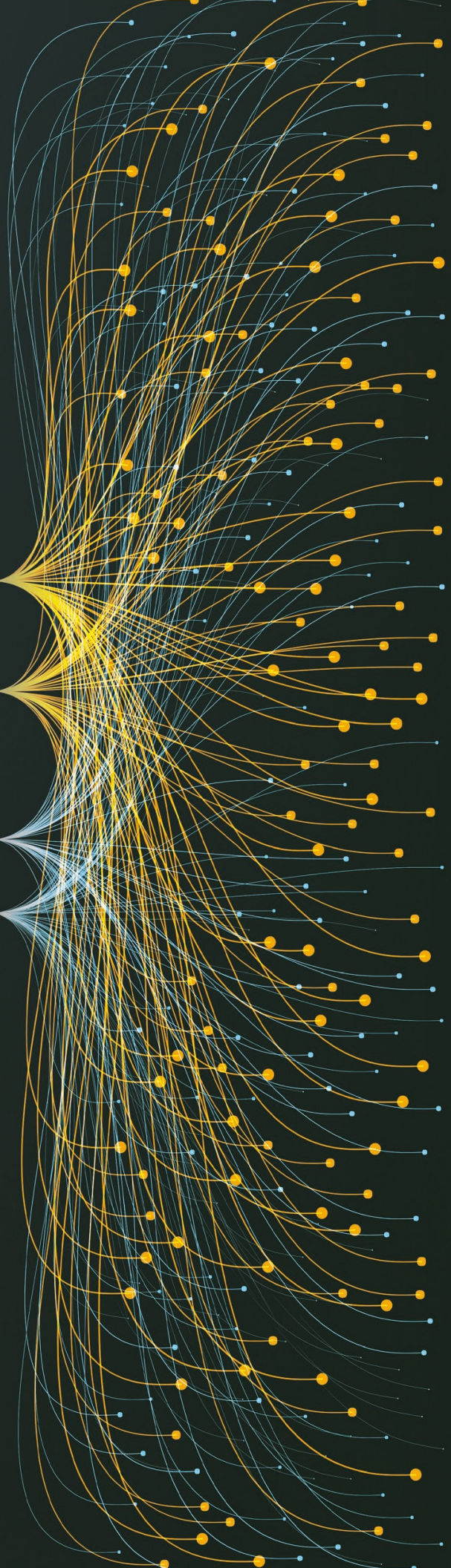
Boost success of your company with Mathematics



BILBAO **15-19** MAY 2017

ESGI {131}

European Study Group with Industry

wp.bcamath.org/esgi131

(bcam)
basque center for applied mathematics

IN COLLABORATION WITH

Bizkaia beaz

[math-in]net
Red Española Matemática-Industria

MI-NET

PARTICIPATING COMPANIES

CIE Automotive

EROSKI

fon

Xpheres
BASKETBALL MANAGEMENT

PROCEEDINGS OF THE
131st EUROPEAN STUDY GROUP WITH INDUSTRY

BCAM – Basque Center for Applied Mathematics
Bilbao, 15–19 May 2017

ESGI {131}

EDITED BY

Ainara Gonzalez, Dae-Jin Lee and Nagore Valle (Basque Center for Applied Mathematics)
Contact: esgi131@bcamath.org

Contents

Preface

The 131st European Study Group with Industry (131 ESGI) was held in Bilbao, Spain, from 15-19th May 2017. It was organized by BCAM - Basque Center for Applied Mathematics, in collaboration with BEAZ (public company of the Provincial Council of Bizkaia), MATH-IN (Spanish Network for Mathematics and Industry) and UPV/EHU (University of the Basque Country).

In addition, ESGI 131 was financially supported by the Basque Government, Severo Ochoa Excellence Accreditation and the Mathematics for Industry Network (MI-NET), COST Action funded project TD1409, which aims to facilitate more effective widespread application of mathematics to all industrial sectors, by encouraging greater interaction between mathematicians and industrialists.

Study Groups with Industry are an internationally recognized method of technology transfer between academia and industry. These one-week long workshops that started in Oxford in 1968 provide an opportunity for engineers and industrial developers to work together with academic mathematicians, students and young professional mathematicians on problems of direct practical interest. Their main objectives are to increase activity in Industrial Mathematics, spread awareness of the benefits of Mathematics, encourage the interaction between researchers from different areas and promote collaborative R&D projects between research groups from the academic sector and companies, addressing problems that can be solved with mathematical models and computational techniques. The Scientific Committee of ESGI 131 selected four problems to work on:

1. Parametric Design by Computational Fluid Dynamics Simulation (CIE Automotive)
2. Improvement of the Contact Center Performance (Eroski)
3. Self-Organized Networks (Fon Labs)
4. Big Data in Sports: Predictive Models for Basketball Player's' Performance (Xpheres Basketball Management)

At the beginning of the week, the 35 participants were divided into groups. Each group worked as a team on one of the problems proposed by the companies mentioned above. On the last day of the workshop, the working groups presented their progress in solving the problems and the recommended approaches. The study cases are assembled in this Study Group Proceedings Report, which provides a formal record of the work for both the industrial and the academic participants. The description of the problems, and the final reports of each working group, as well as a copy of this document, are posted on the website of ESGI 131: https://wp.bcamath.org/esgi131/

The Scientific Committee of ESGI 131 was formed by the following members:

- Elena Akhmatskaya, BCAM – Ikerbasque
- Laureano Escudero, Rey Juan Carlos University
- Luca Gerardo-Giorda, BCAM
- Carlos Gorria, UPV/EHU
- Dae-Jin Lee, BCAM
- Mikel Lezaun, UPV/EHU
- Jose Antonio Lozano, BCAM – UPV/EHU
- Ali Ramezani, BCAM

Challenges description

Challenge 1: Parametric Design by Computational Fluid Dynamics Simulation

DESCRIPTION

In the great majority of industrial problems involving fluid mechanics, turbulence effects must be taken into account. The physics of turbulence is extremely complex since non-linear effects lead to chaotic motion of the fluid, involving very different scales of space and time (multi-scale problem). In order to predict the fluid's mechanical behaviour, it is necessary to resolve the different scales as much as possible. Computational Fluid Dynamics (CFD) simulations of this type typically require considerable computational effort and time (in the order of weeks), for both designing the computational meshes and running the computer calculations.

Design optimization in the automotive industry is an interesting application of these type of numerical simulations; however, because they are currently unfeasible (due to their high computation cost), much faster prediction of the fluid flow is needed in order to run a design of experiments (DOE) with a combination of several parameters and multiple simulations.

OBJECTIVE

Implement effective techniques for accelerating the solution by:

- Identifying regions of the domain where the fluid flow characteristics are more influential for the problem, while saving computation cost by lowering the resolution in regions with lower turbulence intensity;
- Reducing the number of necessary numerical simulations with model reduction techniques (e.g. Proper Orthogonal Decomposition, Fast Fluid Dynamics, Scale-decoupling, etc.);
- Considering other effective techniques proposed by the participants.

Challenge 2: Improvement of the Contact Center Performance

DESCRIPTION

Since 2013, The Eroski Group has been immersed in the SIEC (Integral, Efficient Customer Service) Project, the objectives of which are:

- To provide a multipurpose contact point to internal customers.
- To define and implement the processes, technologies and equipment necessary to carry out that mission.

An important component of this project is the Contact Center, a specialized customer care team which carries out the tasks of contact with the customer, logging of incidents, correct escalation of incidents to specialist teams, and proper case closure.

OBJECTIVE

So that the Contact Center is efficient (i.e., it attends to customers with the appropriate measures according to the established service parameters), and taking into account that said contact center has a two-tiered structure (TIER 1 consisting of agents providing service in the first instance and TIER 2 teams composed of specialists in each area served), we seek a sizing model that allows us to properly manage the flow of incidents. Up to this point, we have only been able to use (ERLANG) sizing methods for TIER 1, having found no model for TIER 2. The proposed improvement project is directly related to optimizing resource allocation, reducing problem resolution time, and streamlining the process between the identification of a problem and the search for solutions. Some of the points and techniques that should be considered are:

- Proposal of a model that describes the operation of the Contact Center service, along with the flowchart indicating the order of stages, priorities, etc.
- Review of the indicators that are currently used to catalogue incidents and their level of resolution, and definition of new measurable variables in the model that are useful for examining solutions to the problem.
- Analysis of the efficiency of processes, software and technologies currently used.
- A study of some aspects of the theory of Operational Research, particularly in network flow models, queuing theory or stochastic programming to formulate a deterministic version of the model that is capable of applying optimization methods.

Challenge 3: Self-Organized Networks

DESCRIPTION

Fon is the world's leading carrier WiFi provider. Pioneers of residential WiFi sharing, we revolutionised carrier WiFi with our technology, creating a globally connected WiFi network. Today, we continue to innovate through two leading business areas. Fon Solutions offers best-in-class WiFi products and services. Our cutting-edge management solutions enable service providers to configure, deliver and operate their own WiFi services. Fon Network aggregates residential and premium carrier WiFi footprints creating one coherent global WiFi network. We facilitate WiFi interconnection between carriers, provide access deals to interested parties, and enable seamless user roaming. Fon's global clients include British Telecom, the Deutsche Telekom Group, SFR, Proximus, KPN, Cosmote, MWEB, SoftBank, Telstra, and Vodafone.

Designed specifically with Communications Service Providers (CSPs) in mind, Fon's cutting-edge WiFi Service Management Solution allows these companies to deliver WiFi services to subscribers and manage them just like cellular and fixed services, in a secure, scalable and flexible way.

WiFi networks are currently one of the main access technologies to the Internet, thanks to their low cost and easy deployment. However, their high density of WiFi access points may impact performance as the deployment is often unmanaged, unplanned, not coordinated in any way and consequently, far from optimal.

When a large number of WiFi hotspots are located within the same coverage area, it is likely that they operate in interfering frequencies with varying power levels. This has a severe impact on user performance due to the medium access mechanism defined in the 802.11 standard (CSMA/CA), whereby each user first listens to the medium and then only transmits if the listened channel is unoccupied.

OBJECTIVE

To develop an intelligent optimization algorithm to coordinate the frequency selection at the back end (for radio resource management purposes), in unmanaged, partially cooperative urban environments where not all the hotspots can be configured. The expected outcome of the algorithm is to:

- Minimize the number of interfering transmitters in the same contention domain, in areas where the spectrum is particularly crowded.
- Provide frequency channel planning at an urban district level.

Challenge 4: Big Data in Sports: Predictive Models for Basketball Player's' Performance

DESCRIPTION

Data analytics in professional sports has experienced rapid growth in recent years [1]. Development of predictive tools and techniques began to better measure both player and team performance. Statistics in basketball, for example, evaluate a player's and/or a team's performance [1,2].

Xpheres Basketball Management is one of the leading basketball player representation agencies in Spain and Europe. A database with men's professional basketball statistics from the last 16 seasons in more than 25 professional leagues and 71 FIBA tournaments has been obtained from **Aryuna**©, a platform that allows performing advanced data analytics of men's professional basketball Statistics. The complete database consists of more than 37,000 games and upwards of 20,000 players.

OBJECTIVE

Based on a database, we aim to:

1. Characterize the performance curve, peak and optimal age in professional men's basketball using performance ratings of players in top European leagues.
2. Determine a rating correction factor for different basketball leagues, which accounts for intra-league and cross-league variability as well as for player characteristics (position, age, player ratings, etc.).
3. Determine which are the most important factors for predicting future outcomes (a successful professional career) of a basketball player.
4. Study statistical models to evaluate the performance of a player based on position, age, skills, league and other characteristics, and their influence in the game.

REFERENCES

[1] https://en.wikipedia.org/wiki/APBRmetrics
[2] https://en.wikipedia.org/wiki/Basketball_statistics

Final reports

Challenge 1: Parametric Design by Computational Fluid Dynamics Simulation

Academic coordinators Ali Ramezani[1] (aramezani@bcamath.org), Laura Saavedra[2] (laura.saavedra@upm.es)
Institution BCAM – Basque Center for Applied Mathematics[1], Universidad Politécnica de Madrid[1,2]

Participants Ferran Brosa[1], Jesua Israel Epequin[2], Nabil Fadai[1], Antonio Zarauz[4]
Institution Oxford University, UK[1], Paris VI University, France[2], University of Almeria[4]

Business coordinator Jon Ezkerra, Gotzon Gabiola
Company CIE Automotive

[UNDISCLOSED CONTENT]

Challenge 2: Improvement of the Contact Center Performance

Academic coordinators Josu Doncel (josu.doncel@ehu.eus), Carlos Gorria (carlos.gorria@ehu.eus), Mikel Lezaun (mikel.lezaun@ehu.eus)
Institution UPV/EHU – University of the Basque Country

Participants Elene Anton[1], Christian Carballo[1], Edurne Iriondo[2]
Institution UPV/EHU – University of the Basque Country[1], University of Zaragoza[2]

Business coordinators José Enrique Rey, Jose Luis Oscoz, Larraitz Tejeria
Company Eroski S. Coop.

ABSTRACT:

The design of an efficient strategy for management of the answer on a multilevel call center distributed by and multi-skill servers is a bit of a challenge. In this type of hierarchic schemes, the first line agents receive a call flux where the waiting time follows an exponential distribution where the dynamics fit with the canonical Erlang-C model. However, the second level agents are skill-specialized and the servers have to deal with many categories of demands and different deadlines to be solved. In this case, the queuing theory doesn't give a successful solution to the problem and routing algorithms and staffing strategies have to be proposed to improve the satisfaction and achievement indicators.

keywords: Multi-level Call Centers, Priority Jobs.

Introduction

The Eroski S. Coop. is a big distribution company operating principally in the North of Spain. The Company relies on the support of a call center service devoted to managing and resolving the incidences that appear in every section of the organization.

The call center is made up of three levels or agents of servers. The first line agents receive all the incidents by phone, email or web form, they label them by categories and subcategories and try to solve (in approximately 10% of the cases) or redirect to the second level. The second level agents are designated in one of the four specialized groups: informatics, commercial, logistics or general/maintenance. Another 10% of the incidences are solved at this point. The third level agents, in general, are external technicians to the Company and they operate on-site at the place where the incidence has been originated.

The customers who use the call-center are all kind of workers, agents, departments, shops, providers, etc, belonging to the human infrastructure of the Company. There are two goals for this challenge. On one hand, it is necessary to plan an optimal scheduling of the human resources available in the second level of the system in order to offer a quality service. On the other hand, in would be useful to find an efficient routing algorithm to channel the incidences from the first level agents to the second level technicians to deal with the priorities of the incidence, the remaining deadline and the skills of the designated agents.

The quality indicators mainly used in this context are the AWT (Accepted Waiting Time) and the SL (Service Level), that is calculated as the proportion of incidences with a response before the established AWT.

The modeling of the dynamics of the commercial call centers has been widely studied in the last decades [3], [7]. Efficient strategies have been proposed for estimating the number of servers working with the aim of guaranteeing short waiting time on the queue and several routing algorithms have been designed for the assignment of tasks to the servers. The dynamics of the flow registered and managed at the first level of the center are efficiently modeled by the Erlang-C pattern under the queuing theory [1]. In practice, the rate of arrivals by unit time μ and the rate of solved or redirected incidences by unit time λ are calculated at any time interval [5], [6]. The formulation allows estimating adequate numbers of servers to be working in order to fulfill the quality rates.

Otherwise, the situation at second level is radically different due to the distribution of the incidences by categories and the assignment to the agents depending on their skills. The new scenario undergoes two main difficulties. First, the disaggregation of incidences contributes to a logical worsening on the statistical properties of the model. Secondly, the duration of the attention paid by the servers follows a very irregular behavior. In some occasions the origin of the incidence is informatic and the solution can be straightforwardly given. In other occasions, the incidence has to be redirected or scaled severaltimes to third level technicians and the tracking has to be made by the second level agent until the customer is informed of the resolution and the incidence is closed.

The complexity of the problem dissuades from using the queuing theory by the difficulty of getting analytical results from the formulation of such irregular statistical distributions. However, the skill-based routing strategies[4], [8] and the grid optimization theory may give rise to efficient algorithms in order to improve the occupation rate of the servers and the proportion of incidences closed on the acceptable time [2].

In the following sections, the problem is dissected in the scenario assumed for the first level agents and for the second level agents of the call center. In the second case, it is proposed a skill-based routing algorithm, where the assignment of the incidences to each agent is weighted dynamically in every unit of time by a factor depending on the skill of the agent, the priority of the task and the time remaining until the deadline for these type of tasks is reached.

Queueing Theory Approach

The mathematical modelling of call centers has a long story and the most important tools that have been used in this context are of queueing theory. In fact, the first level of Eroski's call center is very predicted using one the ERLANG-C model, which is a fundamental model in queueing theory. In Figure 1 we illustrate the ERLANG-C model. As it can be observed, jobs arrive to the queue and, when a server gets idle, it is a job from the queue.

As we said before, the call center that Eroski handles is composed by different levels. Unfortunately, to the best of our knowledge, there is no mathematical model that analyzes the response time of tasks in multi-level call centers. Therefore, in this work we focus our attention in the second level of the call center.

Let us first present the second level of the call center as it is currently implemented. Jobs that arrive to the second level are classified in four different types: informatics, logistics, commercial and maintenance. The workers in this system has a single skills and workers of a given type receive tasks only of their type. Therefore, this system consists of four independent ERLANG-C type systems. There is a vast literature investigating the performance of queueing systems, such as those composed by independent ERLANG-C models. It is known that the latter system can lead to a significant inefficiency. For instance, consider that a informatics type job arrive to the system when the workers of informatics are busy and the rest are idle (and they could therefore not be handled by workers of

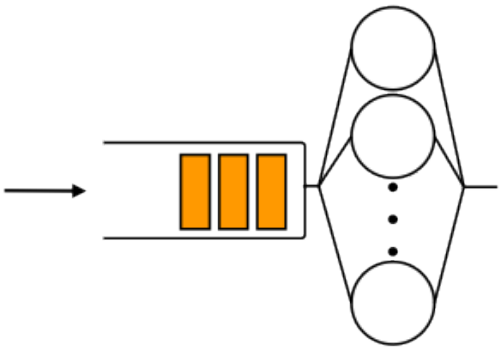


Figure 1: The ERLANG-C model.

Current	New settings
1. Informatics	1. Informatics(H) + Logistics(L)
2. Logistics	2. Logistics(H) + Informatics(L)
3. Commercial	3. Commercial(H) + Maintenance(L)
4. Maintenance	4. Maintenance(H) + Commercial(L)

Table 1: New working groups. (H) means high willingness and (L) low willingness.

other types). It is clear thus that, when the call center is composed by workers with a single skill jobs, its performance can be improved. Following what we have said in the previous paragraph, we design a call center where workers have two skills. Hence, the call center we are presenting is formed by two independent ERLANG-C systems, where in each of them there are workers with two skills, that is, they can handle tasks of two different types. We consider that workers have some preference to handle a task of a given type. For instance, worker of logistics type are able to perform also tasks of maintenance since they have been trained to do it. However, it is clear that they prefer to execute tasks of informatics type.

We also study the load balancing or routing policy in the systems, that is, how to decide which is the task of the queue that is going to be sent to a worker when it gets idle. In the next section we present the algorithm we propose. We believe that its implementation is feasible since it only requires that workers get some training about another skill. Moreover, as our simulations show, the efficiency of the call center gets substantially improved by implementing the system we propose.

Efficient Routing Process

For the new approach related to the routing process, new groups are done for the workers. As we can see in Figure 1, skills are mixed and there are different willingnesses.

Groups that can process the same jobs are different because of the willingnesses or skills resulting from the grouping process, so it is defined

$$s_{ij} : \text{willingness to do job } i \text{ by worker class } j \ .$$

For example, it can be defined $s_{ij} = 0.8$ for H(igh) skill and $s_{ij} = 0.2$ for L(ow) skill.

Notice that in the case of new settings on the model of Figure 1, the problem can be seen as two independent problems, since groups 1 and 2 are independent from groups 3 and 4. Thus, first two groups, the ones involving Informatics and Logistics can be treated without loss of generality in order to develop conclusions.

An arriving job i has the following initial information:

- D_i: deadline time, i.e., maximum amount of time to solve.
- d_i: expected service time for job i.
- t_i: time at which job i arrival happened
- u_i: initial priority level

It is defined a priority function at time t for job i:

$$g_i(t) = \left(\frac{1}{(D_i - d_i) - (t - t_i)} \right)_+ .$$

This function grows while the deadline is closer, taking also into account the time needed to process the job. We need to add the willingness of a worker type j to the job i to the formula:

$$s_{ij}\, g_i(t)$$

For example, there are two groups of Informatics + Logistics, each of them preferring their primary skill, and this property was added to the model. Notice that willingness holds:

- $s_{INF,INF} \geq s_{INF,LOG} = s_{LOG,INF} \leq s_{LOG,LOG}$
- $s_{INF,MAI} = 0$, that is, they do not share knowledge.

The evolving priority for job i to worker j was defined as

$$s_{ij}\, g_i(t).$$

We need to add to the formula the fact that there can be incoming Urgent labeled jobs and force them to be served as fast as possible. Then, it is proposed

$$q_{ij}(t) = s_{ij}(g_i(t) + u_i)$$

With $Q(t) = (q_{ij}(t))$, it is available the information of the actual priority of the remaining jobs, which is used and updated after an event. This is the key of the routing algorithm.

The algorithm for the efficient routing process, based on the defined matrix $Q(t) = (q_{ij}(t))$, is the following:

1. At any time we have the $I \times J$ dynamic size matrix $Q = (q_{ij})$, where I is the number of waiting jobs and J the number of available workers.
2. An available worker j chooses the job

   $$i = \arg\max_{i \in I} q_{ij}$$

   Then delete row(job) i and column(worker) j from Q as there are no longer available.
3. If a new job i arrives at time t_i, we add a new row to Q with the corresponding values $q_{ij}(t_i)$ and update Q to time t_i.
4. If a worker j finishes his job at time t_j, we add a new column to Q based on the formulas $q_{ij}(t_j)$ and update Q to time t_j

Simulation Results

The performance of the proposed solution has been measured by means of simulation. The aim of this study is to test our proposed solution against Eroski's current system. By way of illustration, a simple version of Eroski's Call Center second level has been considered. In our second level, there are two workers and incoming tasks are of two types: *Informatics* and *Logistics*. Worker 1 is mostly skilled in Informatics, but can complete Logistics tasks when needed. In the same way, worker 2 is most skilled in Logistics, but can also do Informatics jobs. A simulated set of incoming tasks has been created, under the assumption that the arrival process is a Poisson one. Table shows the set of tasks arrived in a 10 time units period. Task are classified by priority and topic, and an estimation of their completion time is given.

	priority	topic	completion time
1	No priority	L	4
2	No priority	I	4
3	Priority	I	4
4	No priority	I	2
5	Priority	I	2
6	No priority	L	3
7	Priority	I	3

Table 2: Arrived tasks in a 10 time units period

When entering the system, a priority level q is assigned to each task, for each worker. As defined in the previous section, its value depends on the intrinsic priority of the job and on the skillfulness of each worker. This priority level q increases as the task waits in the queue, and becomes zero when it is taken by a worker. Figure 2 shows the evolution of the priority level for each incoming task. Logistics tasks are marked in blue, whereas Informatics tasks are marked in red.

Those jobs completed by worker 1 have plus symbols, those completed by worker 2 filled circles and uncompleted tasks unfilled ones. The time line in Figure 3 summarizes each worker's activity throughout the considered time period. Each task is numbered as in Table and colored by topic.

Figures 2 and 3 can be interpreted as follows. When the simulation begins, a Logistics task enters the system. This job is chosen by worker 2, as he is the most skilled one in Logistics. At $t = 2$, an Informatics task arrives and worker 1 takes it. In our proposed system, worker number 1 would have taken this task regardless of its topic, as he is free and, to a certain extent, skilled for both topics. At $t = 4$ an Logistics urgent task arrives. Both workers are busy, so it joins the queue. Right after, a non-urgent Informatics tasks arrives and worker 2 completes his job. He must take the urgent task immediately, despite he is more skilled for the Informatics one. Consequently, this task stays in the queue, and its priority increases with time. Another simulation has been carried out in order to imitate the behavior of Eroski's current system, where workers are single-skilled. Table 3 compares both systems' performance.

Our proposed system is able to cope with 6 out of 7 jobs, while Eroski's current system leaves 3 tasks on hold. Moreover, unlike the current system, the proposed solution completes every urgent

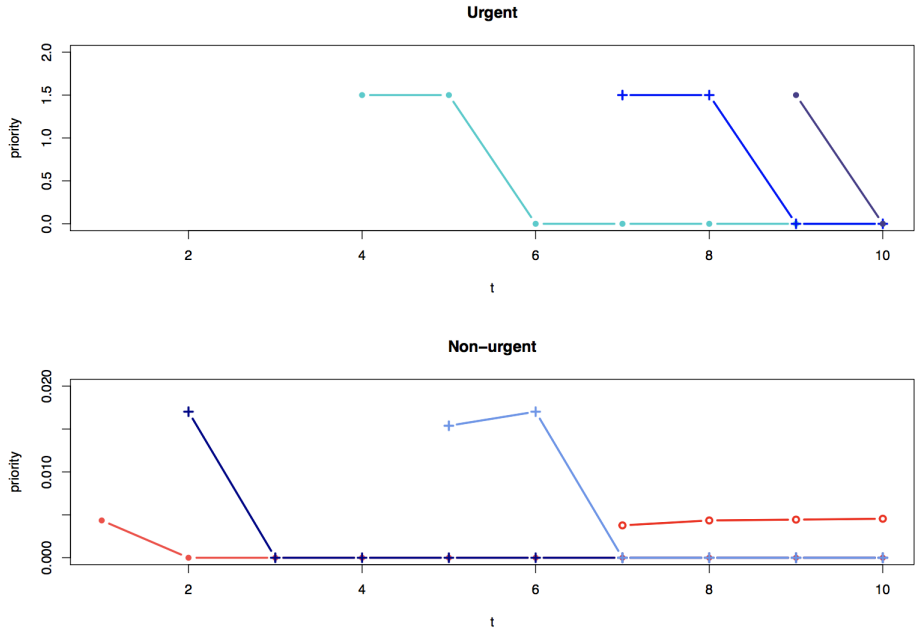


Figure 2: Evolution of the Q matrix

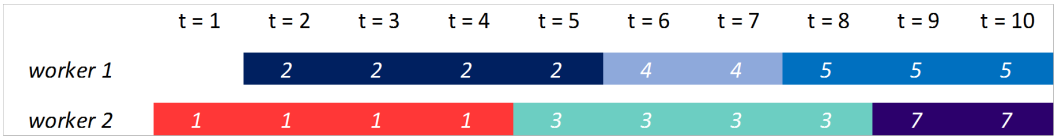


Figure 3: Worker's evolution

	New queuing system	Current queuing system
Taken tasks	6	4
Priority tasks	3/3	1/3
Worker 1	2,4,5	2, 3
	(3/3 main skills)	
Worker 2	1,3,7	1, 6
	(1/3 main skill 2 priority)	

Table 3: Both systems solutions

task. This is an small example that shows how the performance of system can be dramatically improved when two-skilled workers are considered, and an appropriate numerical priority value is assigned to every task.

Conclusions

The main problem analyzed by the working group has been to design an efficient algorithm that guarantees an improvement on the expectation of the occupation rate of the agents of the second level at the same time that increases the rate of response to the incidences on time. In absence of regularity on the flow of calls as well on the response time due to the wide variety of incidences, a skill-based routing has been proposed. The assignment to the servers is weighted dynamically in every unit of time by a factor depending on the skill of the agent, the priority of the task and the time remaining until the deadline for these type of tasks is reached. Some flexibility over the skills of the agents has been taken into account. In fact, any group of agents has been considered as highly qualified in one skill and not so qualified but capable in other skill. The simulations show how this flexibility on the capability of the agents gives rise to an important improvement on the quality rates of the system.

Bibliography

[1] G. Capdehourat, *Análisis y Diseño de Call Centers*. Tech. Rep. 2006.

[2] S.C. Borst, P.F. Seri, *Robust algorithms for sharing agents with multiple skills*. Technical Memorandum BL011212-970912-16TM, Bell Laboratories, Lucent Technologies, Murray Hill NJ. 1997.

[3] G. Koole, *Call Center Optimization*. MG Books. 2000.

[4] R.B. Wallace, W. Whitt, *A Staffing Algorithm for Call Centers with Skill-Based Routing*. Manufacturing & Service Operations Management, 7(4):276 – 294, 2005.

[5] D. Gross, J.F. Shortle, J.M. Thompson, C.M. Harris , *Fundamentals of Queueing Theory*. Wiley. 2008.

[6] M. Singer, P. Donoso, A. Scheller−Wolf, *Una introducción a la teorí a de colas aplicada a la gestión de servicios*. Abante, 11(2):93˘120, 2008.

[7] M. Koole, A. Mandelbaum, *Queueing Models of Call Centers: An Introduction*. Annals of Operations Research, 113:41˘59, 2002.

[8] O. Garnett, A. Mandelbaum, *An introduction to skills-based routing and its operational complexities*. Tech. Rep. 2000, http://iew3.technion.ac.il/serveng.

Challenge 3: Self-Organized Networks

Academic coordinator Javier Del Ser (jdelser@bcamath.org)
Institution BCAM, TECNALIA, UPV/EHU

Participants Aleksandra Stojanova[1], Dusan Bikov[1], Gorka Kobeaga[3], Mirjana Kocaleva[1], Thimjo Koca[4], Thomas Ashley[4], Todor Balabanov[6]
Institutions Goce Delchev University, Macedonia[1], UPV/EHU – University of the Basque Country[2], BCAM – Basque Center for Applied Mathematics, Spain[3], Autonomous University of Barcelona, Spain[4], University of Seville, Spain[4], Bulgarian Academy of Sciences, Bulgary[6].

Business coordinators José Pablo Salvador, David Valerdi
Company Fon Labs

ABSTRACT:

During recent years the number of WiFi networks has experienced rapid growth. The ever-growing number of wireless communications systems have made the optimal assignment of a limited radio frequency spectrum a problem of primary importance. With the common 802.11 wireless technology, few non-overlapping channels are available, and there is no standard mechanism for the access points to dynamically select the channel to be used in order to minimize interference with other access points. This has resulted in a situation where many WiFi networks use default or suboptimal channel assignments, leading to poor performance, and uneven spectrum usage. In this paper, we introduce three different approaches for the frequency assignment problem based on graph coloring that can significantly improve frequency distribution and reduce the number of collisions in the network.

Introduction

WiFi networks are one of the main access technologies to the Internet thanks to their low cost and easy deployment. However, this may impact performance as the deployment is often unmanaged, unplanned and clearly not optimal. Having a large number of WiFi hotspots within the same coverage area, increases the chance that they may operate in interfering frequencies with different power levels. This affects user performance due to the medium access mechanism imposed by the 802.11 standard (CSMA/CA), where each user first listens to the medium and only transmits if it is idle [1].

To access the medium, 802.11 employs a CMS/CA (Carrier Sense Multiple Access with collision avoidance) MAC protocol. Briefly this protocol works the following way: a device that has a packet to transmit, first monitors the channel activity, and if the channel is idle for a predefined period of time, the device will transmit their packet. If the channel is sensed as busy, the device will defer its transmission (for a more detailed explanation on how 802.11 MAC protocol works, please check [7]). As transmissions take place in a shared medium there are two reasons for poor WiFi performance: co-channel interference (CCI) and adjacent channel interference (ACI). The first one, CCI, is when transmissions occur in the same frequency channel. The second source for poor WiFi performance, namely ACI, occurs when transmissions are sent on adjacent or partially overlapping channels. This second effect might not only defer ongoing transmissions, as mentioned above, but also corrupt transmitted frames, leading to an increased number of retransmissions and reduce the WiFi performance of the network. For that reason CCI is preferred over ACI [1].

The literature on frequency assignment problems (FAPs), also called channel assignment problems, has grown quickly over recent years. This is mainly due to the fast implementation of wireless telephone networks (e.g., GSM networks) and satellite communication projects. The renewed interest in other applications like TV broadcasting and military communication problems also inspires new research. Frequency assignment problems arise with application specific characteristics and draw heavily on ideas from graph coloring problems (GCPs) and both are NP-hard problems [5]. Frequency assignment problems are closely related, with the colors now being replaced by frequencies, and edges indicating where interference might occur if the same frequency were used at both ends. In FAPs, numerical labels are used to represent the frequencies, so that more general conditions can be handled than simply the requirement that pairs of frequencies at adjacent vertices be different [5, 6].

Researchers have developed different modeling ideas for handling interference among radio signals or the availability of frequencies, and the optimization criterion. There are different solution methods, which can be divided into two parts. Optimization and lower bounding techniques on the one hand, and heuristic search techniques on the other hand [2, 3, 4, 5, 6].

Our problem was proposed by FON (world's leading carrier WiFi provider), who were looking to develop an intelligent optimization algorithm to coordinate the frequency selection at the back end for RRM purposes (radio resource management), in unmanaged, partially cooperative urban environments where not all the hotspots can be configured.

The expected outcome of the algorithm was to minimize the number of interfering transmitters in the same contention domain, in areas where the spectrum is particularly crowded and provide frequency channel planning at an urban district level.

We proposed three different algorithms and their implementation in Python as a solution to this problem. These algorithms are: Generic algorithm [4], Iterated Local Search [2] and Reinforcement learning algorithm [3].

Iterated Local Search consists of the iterative application of a local search procedure to starting solutions that are obtained from the previous local optimum through a solution perturbation. Local search for the Graph Coloring Problem starts with some initial, infeasible, color assignment and iteratively moves to neighboring solutions, trying to reduce the number of conflicts until a feasible solution is found or a stopping criteria is met. The main goal of ILS is to build a biased randomized walk in the space of the local optima with respect to some local search algorithm. This walk is built by iteratively perturbing a locally optimal solution, applying a local search algorithm to obtain a new locally optimal solution, and finally using an acceptance criterion for deciding from which of these solutions to continue the search. The perturbation must be sufficiently strong to allow the local search to effectively escape from local optima and to explore different solutions, but also weak enough to prevent the algorithm from reducing to a simple random restart algorithm, which is known to typically perform poorly [2].

Reinforcement learning is a learning pattern, which aims to learn optimal actions from a finite set of available actions through continuously interacting with an unknown environment. In contrast to supervised learning techniques, reinforcement learning does not need an experienced agent to show the correct way, but adjusts its future actions based on the obtained feedback signal from the environment. There are three key elements in a RL agent, i.e., states, actions and rewards. At each instant a RL agent observes the current state, and takes an action from the set of its available actions for the current state. Once an action is performed, the RL agent changes to a new state, based on transition probabilities. Correspondingly, a feedback signal is returned to the RL agent to inform it about the quality of its performed action [3].

Reinforcement learning based local search (RLS) combines reinforcement learning techniques with descent-based local search. This approach, also can be used for solving well-known graph coloring problem (GCP) [3].

Summary

This challenge is related to development of an intelligent algorithm to coordinate the configuration of the frequency selection at the backend side for RRM purposes, in unmanaged partially cooperative urban environments where not all the hotspots will be configurable. The algorithm is based on the neighboring hotspots' list, their signal level and their frequency of operation. According the input data, the algorithm generates a frequency channel selection for interference mitigation leading to an optimized deployment of hotspots. It also optimizes the spectrum usage and overall improves user satisfaction given by a better quality of experience and overall higher bandwidth when accessing the network.

Depending on the specific WiFi technology, WiFi devices transmit in two different frequency bands: i) 2.4 GHz, which is typically crowded and ii) 5 GHz, and transmission in each band has different characteristics. Transmissions are configured to take place in a single channel. A WiFi channel has a bandwidth of 20 MHz. The 2.4 GHz band (802.11b/g/n) is divided into 13 channels separated 5 MHz, to that aim only 3 of them are non-overlapping (1, 6, and 11). This frequency channel separation is depicted in Fig. 1, where non-overlapping channels (1, 6 and 11) are highlighted in green.

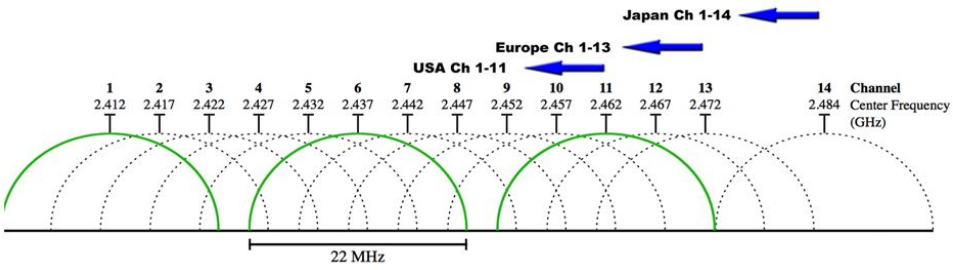


Figure 4: Frequency channel distribution for the 2.4 GHz band.

In the 5GHz band (11a/n/ac) the number of available channels depends on the channel bandwidth, which varies, according to the technology, among 20, 40, 80 or 160 MHz. In this band channels are separated 10 MHZ, thereby there are 24 non-overlapping channels when the hotspots operate with a channel bandwidth of 20 MHz. Broader channels are also exposed to more noise or interference but this band is often less crowded than the 2.4 GHz, which allows for wider channels.

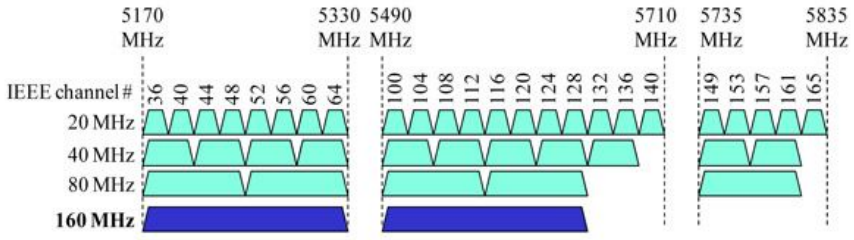


Figure 5: Frequency channel distribution for the 5 GHz band.

Taking everything into account, the objective function used to evaluate the solutions quality has been the following:

$$f(x) = 10 * \log 10(\sum II_i + BI_i)$$

where II_i and BI_i are the inner and the background interferences in Watts of the i-th AP, respectively. The inner interference for the i-th AP can be calculated in the following way:

$$II_i = \sum_{i \in N} \sum_{j \in N \setminus i} x_{i,j} I_{i,j} + \sum_{i \in N} \sum_{j \in N \setminus i} y_{i,j} I^*_{i,j}$$

where

$$I_{i,j} = max_{f \in F} 10^{\frac{s_{i,j,f}}{10}}$$

$$I^*_{i,j} = (1 - \frac{1}{5}|ch(f_i) - ch(f_j)|) \cdot max_{f_1, f_2 \in F} 10^{\frac{s_{i,j,f_1,f_2}}{10}}$$

$s_{i,j,f}$ and s_{i,j,f_i,f_j} are the signal power in decibels and

$$x_{i,j} = \begin{cases} 1, & \text{if the APs } i \text{ and } j \text{ have the same frequency } f \\ 0, & \text{otherwise} \end{cases}$$

$$y_{i,j} = \begin{cases} 1, & \text{if the channels of} i \text{ and } j \text{ satisfy } |ch(f_i) - ch(f_j)| \leq 4 \\ 0, & \text{otherwise} \end{cases}$$

In this work we have use the worst case interference to evaluate the frequency selection, but other variants of the evaluation function should be studied.

Algorithms

Iterated Local Search

The algorithm proposed in this section considers one solution during the search process and two parameters for the tuning of the algorithm. In the first step of the algorithm, a solution is initialized considering only the information of the background interference. For each FON AP we select the frequency that minimizes the interference with the non FON APs. Once we have an initial solution, this is iteratively modified towards better solutions. At each iteration of the algorithm, *param*2 FON

APs are selected and their frequency setting is modified. The number of APs to be changed, *param*2, is an input parameter of the algorithm. The selection of the APs is done based on the interference proportional selection, the APs with a higher interference have a higher probability to be selected.

Algorithm 1 Iterated Local Search

```
for each FON node do
    Select frequency F_i that minimizes the background interference;
end for
Evaluate the total interference I related to the frequency selection F;
it = 0
while it ≤ param1 or I ≠ 0 do
    Update the probability of selecting a node, p_i = I_i / ∑ I_i;
    Select param2 number of nodes using distribution (p_1, ..., p_N);
    for each selected node do
        Define p_f = I_i^f / ∑ I_i^f as the probability of selecting frequency f.
        Calculate the inverse, p_f = (1 − p_f) / ∑ (1 − p_f);
        Sample ones distribution (p_1, ..., p_|F|) to select one frequency f;
        F_i' = f;
    end for
    Evaluate total the interference I' related to the frequency selection F'.
    if I' ≤ I then
        F = F', I = I', it = 0;
    else
        it = it + 1;
    end if
end while
```

For each selected FON AP, i, we evaluate the interference of each possible frequency setting, I_i^f. On the contrary to the initialization, I_i^f considers both the inner interference and the background interference. The channels with lower interference have a higher probability to be selected. After performing the frequency modifications in the APs, we evaluate the goodness of the solution. If the obtained solution is better than the solution that we had at the beginning of the iteration, the modified solution is saved and the count of iterations without improvement is restarted. Otherwise, the modified solution is excluded the count of iterations without improvement is increased by one. The algorithm stops when a certain number of iterations, *param*1, is reached without improvement or when a solution with null interference is found.

Reinforcement Learning Based Local Search

Frequency assignment problems (FAP) can be presented as a subset of graph coloring problem (GCP). As it was described in [3], the graph coloring problem can be attacked with reinforcement learning based local search (RLS). It is a combination of reinforcement learning techniques with local search. In most cases FAP and GCP are NP-hard and that is why they are computationally challenging. If the problems are attacked with exact numerical methods exponential times are expected. The opposite, heuristic and metaheuristic methods are often referred in finding acceptable sub-optimal solution in satisfactory time limit. The negative side of second approach is that solution optimality is not

guaranteed.

Genetic Algorithm

Genetic Algorithms (GA) lie at the core of Evolutionary Computation, a discipline under the wide umbrella of Computational Intelligence that focuses on the exploitations of principles and processes observed in the evolution of species in Nature to construct self-learning methods for optimization and pattern analysis. GA resort to concepts such as genotype inheritance and controlled mutation so as to efficiently explore a search space encoded as chromosomes (individuals or candidate solutions), which undergo probabilistically drive crossover and mutation operators so as to evolve them towards solutions of enhanced optimality for the problem at hand.

A canonical version of the GA utilized for this challenge is described in Algorithm 2. It should be remarked that the crossover and mutation operators included in the above description can be replaced with tailored alternatives more suited to the problem being tackled, possibly by incorporating heuristic information from the problem statement. Likewise, the initialization of the population of individuals can be also performed in a directed manner, e.g. by assigning more probability of initial assignments to those channel(s) with least background interference levels. This, however, can be worked out further beyond the duration of the challenge.

Algorithm 2 Canonical Genetic Algorithm

Initialize a population of P individuals at random from the set of available frequency channels;
Evaluate the total sum interference of the network (fitness) as the sum of the interference received by every FON AP in its selected frequency;
for $it = 0$ to I **do**
  for $p = 0$ to P **do**
    With probability P_c, select two parents from the population based on a selection criterion (e.g. Tournament, Roulette-Wheel), and recombine them by means of an uniform crossover operator;
    With probability P_m, mutate the offspring by using a random mutation strategy;
    Evaluate the total sum interference associated to the produced offspring;
    Add the offspring to the population;
  end for
  Sort the population in increasing order of the sum interference associated to each chromosome;
  Remove the least fit individuals from the extended population, and keep only the best P solutions;
end for

Computational Experiments

In this section, we compare the performance of the GA, RLS and ILS. In order to carry out the comparison, the algorithms have been implemented in Python and run 10 times.

ILS, in all of the runs, was able to find a frequency configuration with null total interference. The mean computational time needed to find such a solution was 285.04 seconds. Results for the GA scheme with $P_c = 0.5$ and $P_m = 0.1$ are depicted in Figure 7 as a boxplot computed over 10 different Monte Carlo realizations. In light of the increased computational complexity of this approach

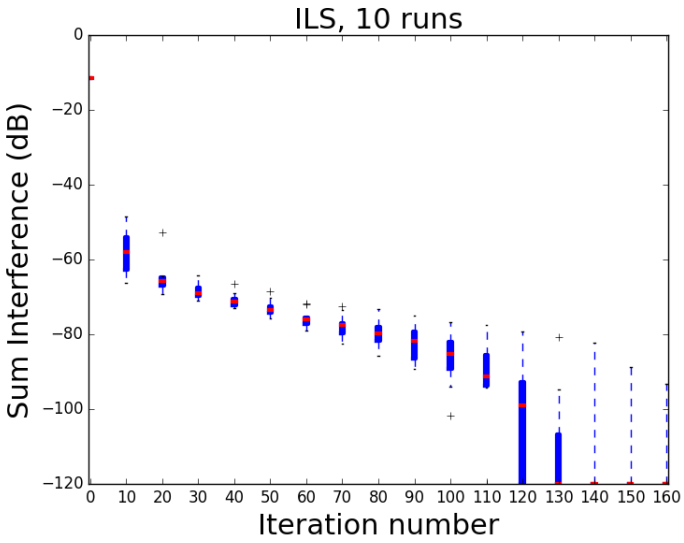


Figure 6: ILS convergence by iterations.

with respect to ILS (330.4 seconds on average) and the lower quality of the produced results our recommendation is to opt for ILS as the practical solution for this problem. It should be noted that GA does not exploit any problem-specific knowledge during its search procedure, hence solutions are provided *blindly* in regards to the particularities of the problem at hand. Incorporation such a knowledge to the definition of GA could be certainly undertaken as future work.

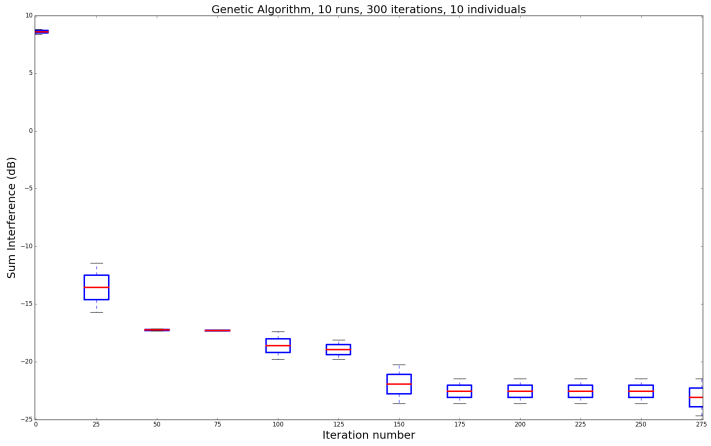


Figure 7: GA convergence by iterations.

At the time of delivery of this report no results were reported by the team for the RLS approach.

Conclusions

As a solution to the problem we proposed Generic algorithm, Iterated Local Search and Reinforcement learning algorithms. By implementing these algorithms in Python and testing them with a given data, we obtained similar results from all three. These three approaches have their advantages and disadvantages but can complement each other. We notice that the heuristic optimization that we used can be very useful for these kind of calculations, but because a large data set was used, these methods can be very time consuming. As further work, we can combine these algorithms or extend the Genetic Algorithm with Iterated Local Search and Reinforcement Learning based Local Search for obtaining better results and more effective solutions. Another further task could be to improve the visualization of the network for better understanding of the obtained results.

Bibliography

[1] J. P. Salvador, Problem Statement: Self Organized Networks, ESGI 131 Challenge Self Organized Networks - proposed by Fon, Bilbao, Spain, 15–19 May (2017).

[2] T. Stutzle, Iterated Local Search - Variable Neighborhood Search, Darmstadt University of Technology Department of Computer Science Intellectics Group, MN Summerschool, Tenerife, Spain (2003).

[3] Y. Zhoua, J.K. Hao, B. Duvala, Reinforcement learning based local search for grouping problems: A case study on graph coloring, Expert Systems with Applications Volume 64, 1 December (2016) 412–422.

[4] G. Colombo, A genetic algorithm for frequency assignment with problem decomposition, Journal International Journal of Mobile Network Design and Innovation archive, Volume 1 Issue 2, September (2006) 102–112.

[5] N. Karaoglu, B. Manderick, FAPSTER - A Genetic Algorithm for Frequency Assignment Problem, onference: Genetic and Evolutionary Computation Conference - GECCO, January (2005).

[6] I. Karen, Aardal,Stan P.M. van Hoesel, Arie M.C.A. Koster, C. Mannino, A. Sassano, Models and solution techniques for frequency assignment problems, Springer Science+Business Media, LLC (2007) 79–129.

[7] B. Kaufmann, F. Baccelli, A. Chaintreau, V. Mhatre, K. Papagiannaki and C. Diot, Measurement-Based Self Organization of Interfering 802.11 Wireless Access Networks, *In IEEE INFOCOM, Anchorage, Alaska* (2007) 1451–1459.

Challenge 4: Big Data in Sports: Predictive Models for Basketball Player's Performance

Academic coordinators Dae-Jin Lee[1] (dlee@bcamath.org), Garritt L. Page[1,2].
Institution BCAM – Basque Center for Applied Mathematics[1], Brigham Young University, USA[2]

Participants Amaia Abanda Elustondo[1], Bruno Flores Barrio[2], Silvia García de Garayo Díaz[3], Manuel Higueras Hernáez[1], Amaia Iparragirre Letamendia[1,3], Mariam Kamal[3], Gorka Labata Lezaun[4], Roi Naveiro Flores[5], Argyrios Petras[1], Simón Rodríguez Santana[5], Quan Wu[3].
Institutions BCAM – Basque Center for Applied Mathematics, Spain[1], University of La Rioja, Spain[2], UPV/EHU – University of the Basque Country[3], University of Zaragoza, Spain[4], ICMAT – Instituto de Ciencias Matemáticas, Spain[5]

Business coordinators Pedro Barrera, Igor Crespo, Oscar Garrido
Company Xpheres Basketball Management.

ABSTRACT:

Data analytics in professional sports has experienced rapid growth in recent years. Development of predictive tools and techniques began to better measure both player and team performance. Statistics in basketball, for example, evaluate a player's and/or a team's performance.

Aryuna© is a platform that allows performing advanced data analytics of men's professional basketball statistics of the last 16 seasons in more than 25 professional leagues and 71 FIBA tournaments. The challenge consisted of the next four goals:

- Characterize the performance curve, peak and optimal age in professional men's basketball using performance ratings of players in top basketball leagues. See [3, 6]
- Determine a rating correction factor for different basketball leagues, which accounts for intra-league and cross-league variability as well as for player characteristics (position, age, player ratings, etc.). See [1]
- Determine which are the most important factors for predicting future outcomes (a successful professional career) of a basketball player. See [2, 6]
- Study statistical models to evaluate the performance of a player based on position, age, skills, league and other characteristics, and their influence in the game. See [5]

Data description and objectives

A database containing a total of 44 variables of 5227 professional basketball players during seasons 2000-2015, and six competitions (Euroliga, Eurocup, ACB, Argentina, ABA, ProA) was provided for the challenge. The participants of this challenge worked in small groups and used the Open Source Software `R` to perform the Statistical analysis and fit statistical models.

Some important concepts:

- **Variables with suffix 'X100Possessions':** are the statistics produced by the player per 100 Team's possessions, e.g. *'Ptsx100Possessions = 20,5'* means the player scores 20,5 points per 100 Team's possessions.
- **Possession:** In basketball, possessions are defined as the time a team gains offensive possession of the ball until it scores, loses the ball or commits a violation or foul.
- **Usage%:** Usage percentage is an estimate of the percentage of team plays used by a player while he was on the floor. The formula is :

  *100 * ((FieldGoals Attempts + (0,44*FreeThrows Attempts) + Turnovers) * Team minutes) / (Player minutes * Team FieldGoals Attempts + 0,44 * Team FreeThrows Attempts + Team Turnovers)*

- **Free throw:** are unopposed attempts to score points from a restricted area on the court (the free throw line; informally known as the foul line or the charity stripe), and are generally awarded after a foul on the shooter by the opposing team.
- **Field Goals:** it refers to a basket scored on any shot or tap other than a free throw, worth two or three points depending on the distance of the attempt from the basket.
- **BPM (Box Plus Minus):** is a box score-based metric for evaluating basketball players' quality and contribution to the team. Visit for details: http://www.basketball-reference.com/about/bpm.html

There are also some definitions that measure the player performance (called *metrics*).

- **EOPx40M(Efficient Offensive Production per 40 Min):** To calculate EOPx40M we need to get OE (Offensive Efficiency).
- **OE (Offensive Efficiency):** Metric that measures the quality of offensive production. An OE of 1.0 correspond to 100 percent efficiency. OE is the total number of successful offensive possessions the player was directly involved in divided by that player's total number of potential ends of possessions. OE formula:

  (FieldGoalsMade + Assists) / (FieldGoalsAttempts - OffensiveReb + Assists + Turnovers)

In order to compute **EOPx40M**, first we calculate EOP, this metric measures the offensive production with a measure of efficiency (OE), it uses points and assists. To use assists in the formula, it needs to know the value of an assist relative to a point scored, the author consider an assist meaningful if the assist led to a basket at the rim. By average about 38% of assists led to a basket at the rim, if a player had 100 assists, he created 2*0,38*100= 76 points, so 1 assists = 0,76 points. EOP Formula:

$$EOP = (0{,}76 * Assists + Points) * OE$$

$$EOPx40M = (EOP / Player\ seconds\ played) * (40 * 60)$$

Remark: OE y EOP are metrics created by Stephen M. Shea in his book Basketball Analytics [8].

Based on the information obtained from Aryuna, the team assigned to this challenge worked in small groups and focused in 4 main goals:

1. Study the performance curve, peak and optimal age of a basketball player in top European leagues.
2. Player's performance and their influence in the game.
3. Rating correction factor for different basketball leagues.
4. Which factors predicts a successful professional career?

Objective 1: Performance of players by age and experience

In order to analyse the performance curves, we chose the variable EOP per 40 minutes (EOPx40M) for each player in one season and one competition. We fitted a mixed-effects model with a quadratic fixed effect for the age of the player interacting with the position. There is also a random effect for each player, to account the variability of each individual. Thus the model is

$$\text{EOPx40M} = \beta_0 \text{Position} + \beta_1 \text{Position:Age} + \beta_2 \text{Position:Age}^2 + u_{\text{Player}} + \varepsilon, \tag{1}$$

where $u_{\text{Player}} \sim N(0, \sigma_{\text{Player}})$ is a random effect per player and ε is the error term, i.e. $\varepsilon \sim N(0, \sigma)$.

Registers of players whose minutes played at the season-competition are lower or equal to 50 minutes are discarded. The data for this model include 10,712 observations for 3,743 different players.

The estimation of the standard deviation of the random effect is $\hat{\sigma}_{\text{Player}} = 2.08$ with p-value < 0.0001, indicating a significant variability for the players.

The fixed effects of this model produce three performance curves, one for each position. The peaks for each position are calculated by calculating their maximums. Let the curve of performance for position p be

$$y = \beta_{0,p} + \beta_{1,p} x + \beta_{2,p} x^2,$$

the derivative is

$$y' = \beta_{1,p} + 2\beta_{2,p} x,$$

and the maximum of the curve (if $\beta_{2,p} < 0$, because $y'' = 2\beta_{2,p}$) is found for the value which solves the derivative equals to 0,

$$\beta_{1,p} + 2\beta_{2,p} x_{\max} = 0 \Rightarrow x_{\max} = -\frac{\beta_{1,p}}{2\beta_{2,p}}.$$

For each position p, $-\hat{\beta}_{1,p}/(2\hat{\beta}_{2,p})$ is the peak performance age and by the δ-method the 95% confidence intervals bound limits are

$$-\frac{\hat{\beta}_{1,p}}{2\hat{\beta}_{2,p}} \pm 1.96 \cdot \left(-\frac{1}{2\hat{\beta}_{2,p}}, \frac{\hat{\beta}_{1,p}}{2\hat{\beta}_{2,p}^2} \right) \cdot \Sigma_{12,p} \cdot \left(-\frac{1}{2\hat{\beta}_{2,p}}, \frac{\hat{\beta}_{1,p}}{2\hat{\beta}_{2,p}^2} \right)^{\mathrm{T}}$$

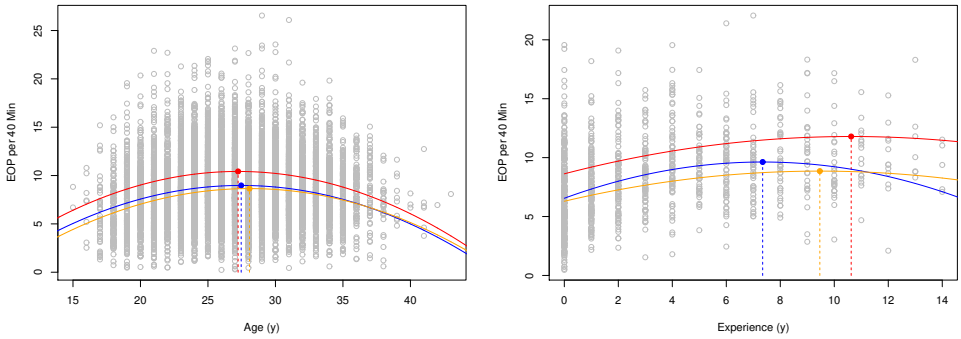


Figure 8: Performance curves by age (left) and experience (right). The grey dots are the observed performances. The solid lines are the performance curves for centers, forwards and guards, and their respective peaks are represented by the solid dots and the dashed lines.

where $\Sigma_{12,p}$ is the variance covariance matrix of $\beta_{1,p}$ and $\beta_{2,p}$. Peak performance ages (and their 95% confidence intervals) by position are:

- Center: 27.23 (26.48, 27.98) years;
- Forward: 27.46 (27.00, 27.92) years;
- Guard: 28.08 (27.57, 28.59) years.

These peaks and their statistical uncertainties are between 26 and 29, which are similar to those for the NBA, 27-30, as shown in [6].

Analogously, curves of performance by the experience of the player interacting with the position have been calculated, with a random effect for each player. As there is no information of the debut season of the registered players, it is assumed that new players from the 2001 season younger than 20 are in their first season. The model has the same structure changing the age of the player by the his experience (time since his debut season). Registers of players whose time contributed in the season-competition is lower or equal to 50 minutes are discarded. The data in this model are 986 observations for 283 different players. Peak performance experiences (and their 95% confidence intervals) by position are:

- Center: 10.63 (2.37, 18.88) years;
- Forward: 7.35 (5.76, 8.93) years;
- Guard: 9.46 (6.38, 12.55) years.

Peak performance experiences are between 7 and 11 years, which are similar to those for the NBA (6-8 years, see [6]). The 95% confidence intervals are huge (mainly for the center players) because of the lack of information (121 centers, 364 forwards and 501 guards). Information about the debut season for each player would give bigger dataset to be analysed in this model. Figure 1 shows the plot of the fixed effect curves by position for both models.

Extraction of Performance Patterns

$$
\begin{array}{c}
 & \begin{array}{cccccc} AGE_{17} & AGE_{18} & \cdots & AGE_{42} & AGE_{43} \end{array} \\
\begin{array}{c} P_1 \\ P_2 \\ P_3 \\ \cdots \\ P_n \end{array} &
\begin{pmatrix}
\emptyset & \emptyset & \cdots & EOP_{P_1,42} & EOP_{P_1,43} \\
EOP_{P_2,17} & EOP_{P_2,18} & \cdots & \emptyset & \emptyset \\
\emptyset & \emptyset & \cdots & EOP_{P_3,42} & EOP_{P_3,43} \\
 & & \cdots & & \\
 & & & &
\end{pmatrix}
\end{array}
$$

Figure 9: New Performance Matrix

PlayerID	17	18	19	20	21	22	23	24	25	26	27	28	29
118	NA	2.020	NA	8.87	11.80	9.640	9.440	8.160	9.940	11.890	15.510	11.740	NA
121	NA	NA	NA	12.13	10.58	11.470	10.080	13.670	14.420	13.500	NA	13.000	14.130
276	NA	10.590	11.840	12.76	14.03	19.550	NA	NA	NA	16.160	12.580	13.710	10.840
280	NA	NA	NA	15.55	13.98	16.250	14.340	12.050	12.070	13.810	17.150	17.890	12.560
284	NA	8.890	NA	9.52	10.71	13.560	10.710	12.130	15.550	13.660	12.480	13.550	NA
325	NA	11.860	7.640	9.31	8.24	8.950	12.190	9.680	7.580	8.250	7.850	7.780	6.790
326	NA	NA	NA	5.06	10.18	10.520	8.750	11.230	9.000	7.470	9.700	7.740	10.190
329	NA	9.010	3.330	9.38	19.55	9.250	10.230	9.540	21.400	12.630	10.700	11.540	11.370
485	NA	NA	NA	NA	9.96	9.100	8.860	8.800	10.080	9.350	8.980	10.110	8.380
486	NA	NA	NA	NA	7.59	9.980	10.650	12.890	10.320	6.625	9.710	6.850	10.840
489	NA	NA	NA	9.60	8.84	NA	NA	NA	11.010	10.070	11.120	9.560	NA
559	NA	NA	NA	NA	NA	NA	NA	11.140	8.980	11.090	10.060	8.880	12.270
560	NA	NA	NA	NA	NA	NA	NA	NA	10.710	9.040	12.560	9.030	11.110
721	NA	NA	NA	NA	NA	NA	NA	NA	9.430	10.720	7.070	10.600	9.680
830	NA	NA	NA	13.60	NA	8.510	10.570	10.925	12.540	13.540	9.490	16.440	10.690
837	NA	NA	NA	NA	6.36	8.480	5.670	NA	12.110	8.020	11.400	11.390	9.130

Figure 10: Example of Performance Matrix of ACB

In order to deal with individual performance trajectories, we split the database according to different competitions and once a competition was selected, we redimensionalized the database as follows: each row d_i represents a player who played in this competition and each column d_j refers to an age. Each element d_{ij} of the new matrix represents the EOP performance index of the player i by the age j.

First of all, it should be mentioned that there are quite a few gaps or undetermined values in the performance matrix due to the lack of data or the restriction to one competition. For example, if a player has been in the ACB for 5 season, then he moved to the NBA for 4 season and finally he came up for 3 season to the ACB, there would be a 4 years long gap in the ACB-Performance matrix of this player.

Therefore, each row on the matrix corresponds to the performance curve of a player in one competition. In Figure 3 some examples of ACB player's performance curves are shown. We only have the points so we tried to do a regression (find the curve that describes those points) with polynomials approximations of different grades.

Nevertheless, this approximations were not good enough and we wanted to go further. If one looks at A. Hervelle and C. Jimenez player's performance curves (the real data, the points in Figure 3), it can be seen that their pattern is very similar. Consequently, we noticed that it make sense to exists different patterns on the performance curves.

In order to extract meaningful patterns, we took advantage of the fact that this performance curve are, actually, time series. That is, each value has been taken by a specific age and the values are, some kind, sorted. In other words, the values on an individual performance curve can not be considered independently, because they are correlated to each other and this is the main characteristic of the time series. Once this contemplated, the goal is to group the performance curve by their patterns, for which a clustering method was used. Clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other clusters, so the aim is to find groups of patterns in the performance curves.

With the intention of clustering the performance curves, a similarity measure between series has to be chosen. Since our interest focuses on the shape of the performance curve, a distance between curve that deals with time warping and shifting was used, namely, Dynamic Time Warping (DTW). DTW is a well-known technique to find an optimal alignment between two given (time-dependent)

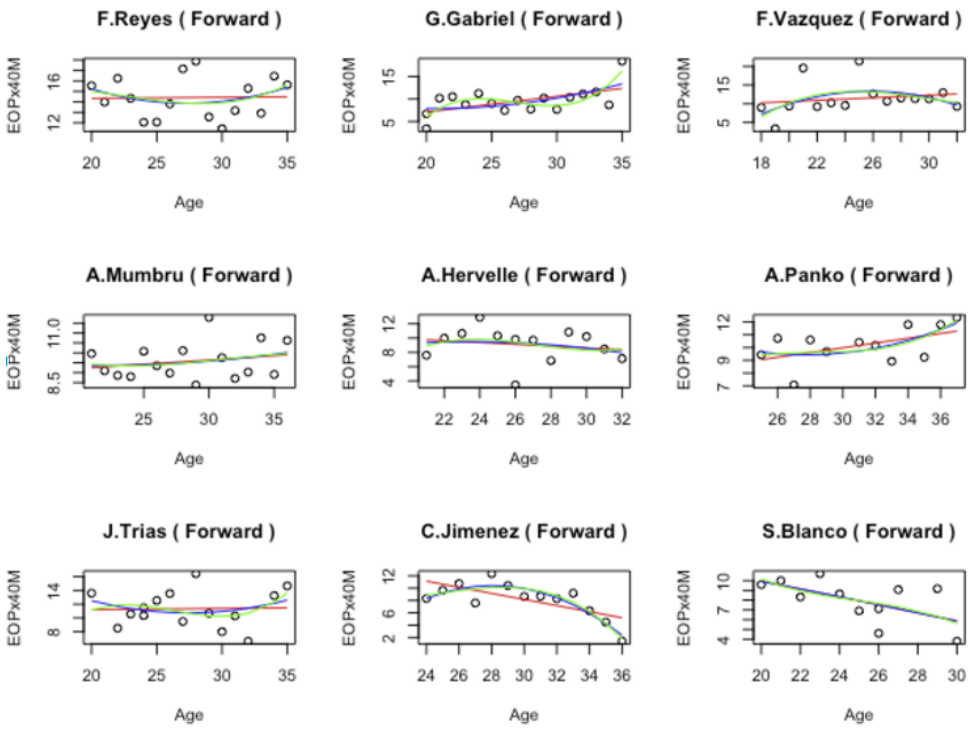


Figure 11: Some performance curves

sequences under certain restrictions. It is able to deal with local warping and shifting by searching the optimal alignment between two series that minimizes the distance, so if two series have the same shape but they are out of phase, Dynamic Time Warping will align them and compute the minimal distance between all possible alignments. Figure 5, illustrates the DTW idea.

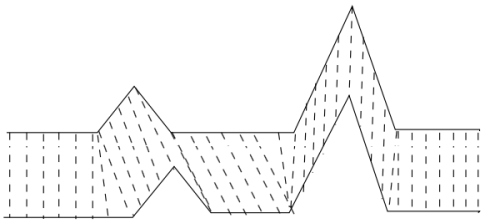


Figure 12: Dynamic Time Warping

Therefore, a hierarchical cluster analysis with Dynamic Time Warping was carried out, with the resulting agglomeration dendogram shown in Figure 6. It illustrates the arrangement of the clusters based on the Dynamic Time Warping distance between performance curves, distinguishing 2 or maybe 3 main clusters. We chose to split de database in 3 clusters, but the same analysis could be made for 2 clusters.

The meaning of this partition is that the cluster analysis show that, based on the similarity of

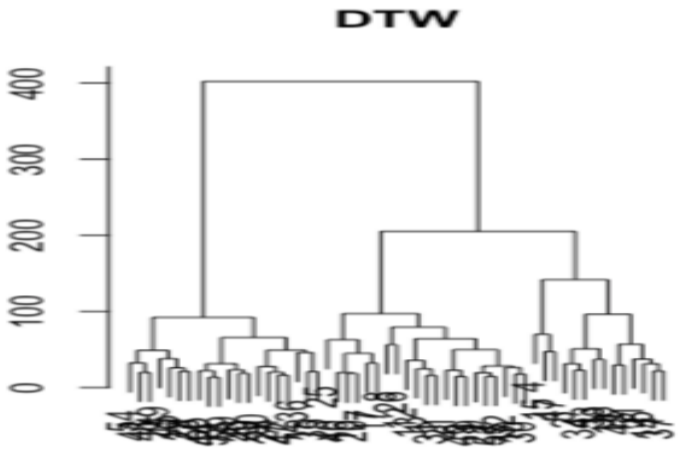


Figure 13: Cluster Dendogram

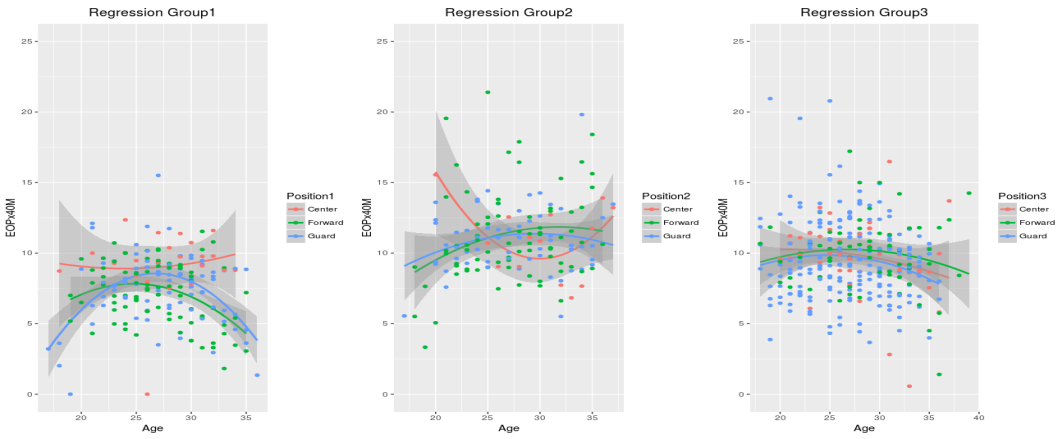


Figure 14: Regression of Clusters by Position (ACB)

the shape of the performance curves, there are 3 main groups. Once the three clusters are divided, we made a regression for each position within each group. The resulting regressions for ACB performance curves are shown in Figure 7. It can be seen the partition on three clusters (each plot shows one cluster) and, in fact, the different patterns for each cluster.

In the first one, the pattern of the performance curve is very similar for Forwards and Guards. Actually, the performance pattern starts with a relatively low EOP in early ages, the players enhance their EOP quickly in the first 5-10 years of competition and reach their performance peak by the age of 25 for Forwards and 27-28 for Guards. After the peak, this kind of players tend to worsen the EOP. It has to be mentioned that for Centers this analysis is not meaningful, due to the lack of data (the error interval coloured by gray is too wide to conclude anything).

The second kind of player (plot in the middle), has a very different pattern of performance. The players in this cluster start with a quite high EOP from a very early ages (as high as the maximum of

the player in the first cluster) and they tend to improve, or increase their EOP, over almost their entire career. The performance peak of this kind of player is reached at the age of 31 for Forwards and 29 for Guards. As happened before, there is not enough information to conclude anything for Centers.

The last kind of player starts very similar to the second one, with a quite high EOP in the first years of competition, but instead of keep improving the performance over the years, they reach a peak and start getting worse. For this kind of player, depending on the position, the peak is reached at 27 for Forwards, 24 for Guards. Again, the lack of data does not allow to conclude anything for Centers.

Objective 2: Player's performance and their influence in the game

For this goal, we considered as player's performance measure the shot precision. Figure 8 shows a boxplot per Competition (for all the seasons) with the kernel density estimator on each side (this is known as *violin plot*). It is shown that the shot precision in all competitions look very similar in 2 points shots (top plots) and very symmetric around their respective averages. The distribution of the shoots for 3 points and free throws are very skewed.

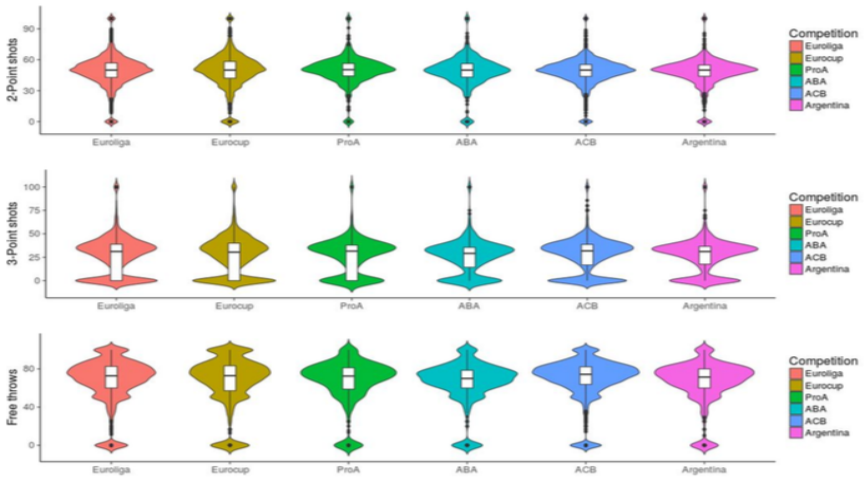


Figure 15: Shot precision by Competition and type of shoot (free, 2 points and points).

Figure 9 shows by player position and age, the increase in % shots in Argentina and Euroleague. There seems to be an increasing percentage in free throws in all the positions. For the rest of the points (2 and 3 points throws) the increase is not so evident due to the high variability of the data.

The conclusions of this goal are not very strong: i) most results are (still) inconclusive due to the need of long-career-players? data; ii) in some rare cases it can be shown that the precision does experience minor changes; and iii) some leagues need more data to obtain better results (if any at all) and finally iv) the nature of each league might be intrinsically different.

Objective 3: Rating correction factor for different basketball leagues

Our second goal was to try to find a conversion factor between the 6 different leagues. As each league has its own characteristics, (some might be more offensive whereas others may use more

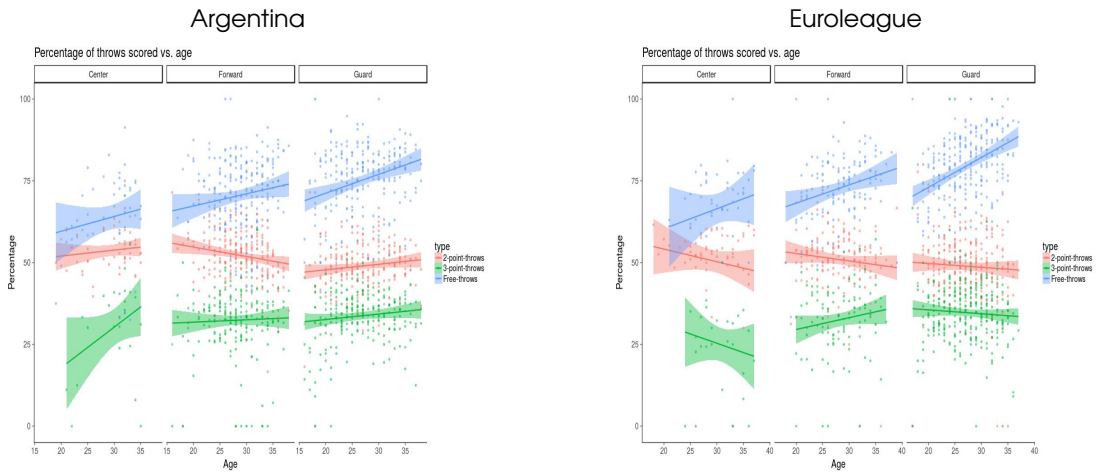


Figure 16: Increase in % shots in Argentina and Euroleague by player position, type of shoot and age.

3 points shots, for example) Xpheres is in the need of finding a tool to transform the performance obtained by a player in one league into the performance in another league. This would let them compare different players from different leagues to find the best ones to represent. Interpretation of the results will be left to experts, so we will just focus on the data.

The biggest problem here is to find differences between leagues. Due to the heterogeneity of players, all leagues look similar, as it can be seen in Figure 10. This happens not only for the EOP coefficient but for any other one too. It cannot be seen any difference in terms of mean, variance of distribution that cannot be attributed to noise in the observations.

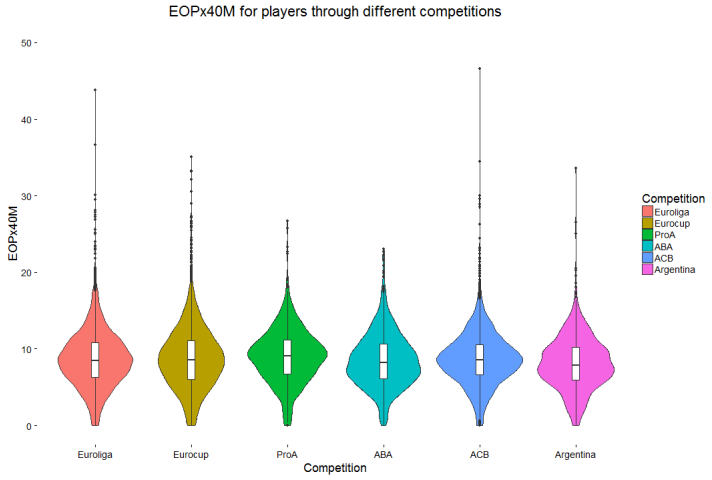


Figure 17: Violin plots obtained for de EOP factor dividing by league.

However, by computing ANOVA (*Analysis Of Variance*) models, we can get some differences for some particular factors and competitions. Using the Euroleague as the base league to compare with, we can see how some means are further than the sum of the standard deviation of both, and all with

small p-values, what would imply that there are differences between those means (see Figure 11).

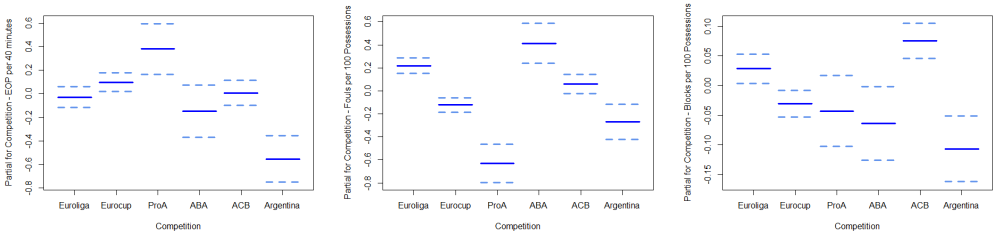


Figure 18: Coefficients plots obtained for EOP (left), Fouls $\times 100$ possessions (center) and Blocks per hundred possessions(right) ANOVA models. The intercept is not taken into account. Continuous lines represent means whereas discontinuous lines represent mean plus or minus standard deviation.

As performance is an abstract concept, many different measures can be built. Indeed, EOP is a measure of offensive performance generated by combination of some basic variables so it might not consider players with high block skills. To consider all different measures, what we propose to do is to predict over those basic variables, so that when a different measure is used, the agent just has to apply the formula of the measure to the predictions of the basic variables it uses.

A complete and very visual table with the differences between means for each basic variable and taking into account the p-values of the ANOVA model can be generated so it will help the agent (see Figure 12).

	Points x 100 Possessions			3 Points Scored x 100 Possessions			3 Points Attempted x 100 Possessions			Free Throws Scored x 100 Possessions			Free Throws Attempted x 100 Possessions			Offensive Rebounds x 100 Possessions		
	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value
Intercept	19,97	0,1	2,00E-16	1,79	0,02	2,00E-16	5,32	0,06	2,00E-16	3,95	0,03	2,00E-16	5,52	0,04	2,00E-16	2,61	0,03	2,00E-16
EUROCUP	-0,06	0,14	0,6795	0,08	0,03	0,00844	0,22	0,08	0,0048	-0,12	0,05	0,0107	-0,18	0,06	0,002924	-0,09	0,04	0,0175
ProA	-0,02	0,23	0,9335	0,07	0,05	0,16838	0,33	0,13	0,0132	-0,67	0,08	2,00E-16	-0,85	0,1	2,00E-16	0,11	0,07	0,1165
ABA	-0,48	0,24	0,0449	-0,05	0,06	0,35107	0,32	0,14	0,0206	-0,17	0,08	0,0278	-0,1	0,1	0,34	-0,07	0,07	0,9232
ACB	0,48	0,15	0,0016	0,15	0,035	1,92E-05	0,39	0,09	5,38E-06	0	0,05	0,97	-0,1	0,07	0,1325	0,09	0,04	0,0346
Argentina	-0,38	0,22	0,0798	0,05	0,05	0,33844	0,51	0,12	4,07E-05	-0,34	0,07	2,40E-04	-0,34	0,1	0,000408	-0,26	0,06	3,61E-05

	Defensive Rebounds x 100 Possessions			Steals x 100 Possessions			Turnovers x 100 Possessions			Assists x 100 Possessions			Blocks x 100 Possessions			Fouls x 100 Possessions		
	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value	Beta	St. Error	p-value
Intercept	5,95	0,04	2,00E-16	2,03	0,02	2,00E-16	3,79	0,026	2,00E-16	3,47	0,04	2,00E-16	0,74	0,01	2,00E-16	6,99	0,04	2,00E-16
EUROCUP	-0,04	0,05	0,41815	0,01	0,02	0,562	0,05	0,03	0,128	0,14	0,05	0,00509	-0,06	0,02	0,00298	-0,34	0,06	7,65E-10
ProA	0,32	0,09	0,00068	-0,06	0,04	0,116	-0,06	0,06	0,314	0,66	0,08	1,18E-14	-0,07	0,03417	0,0367	-0,85	0,1	2,00E-16
ABA	0,12	0,1	0,20468	-0,17	0,04	5,10E-05	0,08	0,06	0,182	0,45	0,09	2,77E-07	-0,09	0,04	0,00868	0,19	0,1	0,04665
ACB	-0,07	0,06	0,26799	0,11	0,03	3,91E-05	0,02	0,04	0,631	-0,04	0,06	0,42644	0,05	0,02	0,04	-0,16	0,06	0,00995
Argentina	0,5	0,09	1,20E-08	-0,26	0,04	3,36E-12	-0,44	0,06	7,26E-15	-0,64	0,08	1,72E-15	-0,14	0,03	2,58E-05	-0,49	0,09	4,96E-08

Figure 19: Coefficients of the ANOVA model for each of the basic variables. Darker colours imply smaller p-values ranking from smaller than 0.001 to smaller to 0.01, to 0.05 and to 0.1.

To be more accurate, we can divide players into different categories. A good start could be to divide by positions as it is not the same the number of 3 points shots a center can score than this number of scores for a forward or guard. We used an ANOVA model with interactions where now dummies variables for belonging to a league and playing in a certain position are added.

Taking into account all player characteristics doesn't have a big computational impact. Although there are just 3 variables to characterize players (age, height and position), some of them can take too many values. Instead of giving a conversion factor for each value, that would give an enormous amount of different players, we use a linear model tree. Since ANOVA models are linear models

with factors, what we are really doing is to fit an ANOVA model at each leave of the tree. Different values for the minimum number of individuals in each leave allow to adjust the number of types of players.

Following the linear model approach, we performed a linear regression tree, allowing for a recursive partition of the variables. Figure 13 shows the obtained tree.

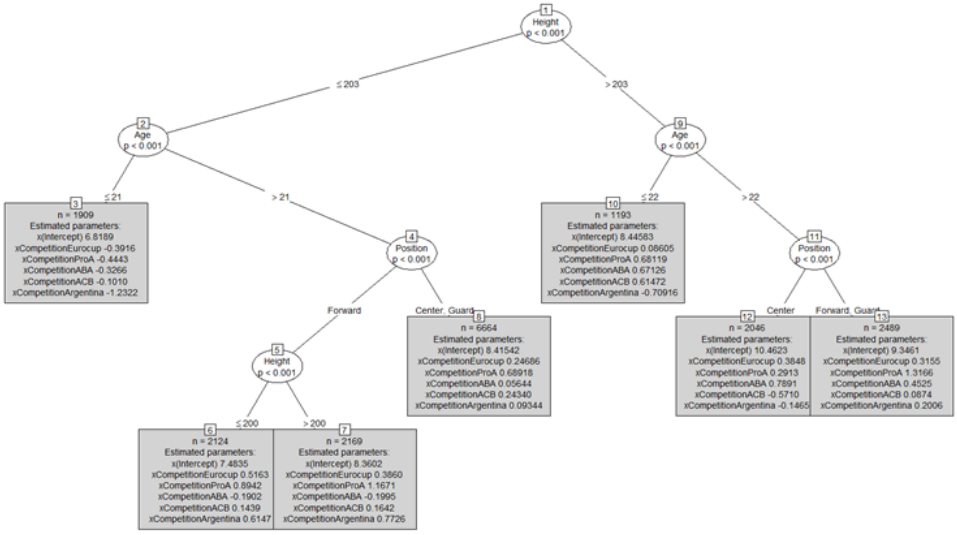


Figure 20: Example of linear model tree to predict the EOP variable adjusted to produce 7 types of players.

The benefits of this last model are that produces good fitting of the data and meantime gives information of the classes of players. All this rules of transformation can be traduced into factors by adding the coefficient in the ANOVA model to the intercept and dividing by the intercept. Thus, a player with a coefficient of 9 in the ACB in a category in which the mean of the Euroleague is 10 and the coefficient of the ANOVA model for the ACB is -0.5 would get a coefficient of:

$$9 \cdot \frac{10}{10-0.5} = 9 \cdot 1.05 = 9.4$$

Other ways to obtain this correction factor could be to just consider players that played at two leagues at the same time (one of them should be Euroleague or Eurocup), but we would not have a conversion factor for Argentina league. We could also consider players who were transferred to other leagues. We could observe the level of a player in one league just before leaving and compare it with the level in the other league a few years later, when we think he got used to the new league.

Objective 4: Which factors predicts a successful professional career?

This is the most challenging task proposed by Xpheres. In fact, from our view there is no relevant information in the provided database to quantify a successful professional career. From a basketball point of view, 'success' could be considered as winning a MVP (most valuable player) trophy at least 3 times in a row during a professional career, signing a contract of several millions of $ etc ... However, we tried to answer the following question: "how can we predict success?".

Any statistical model has to be interpretable and have a strong predictive power. Firstly, we considered the variable *EOPx40M* and transformed this variables in a dichotomous variables (0=Fail and 1=Success). Note that this partition is completely subjective, and was performed only to obtain a measure of success. We performed a popular machine learning technique in order to perform a feature (or variable) selection, called *random forest*. Figure 14, shows the number of most important variables for prediction selected by the random forests, i.e. there are 13 out of the 44 variables that are the most relevant variables to predict a successful player based on the transformation we proposed.

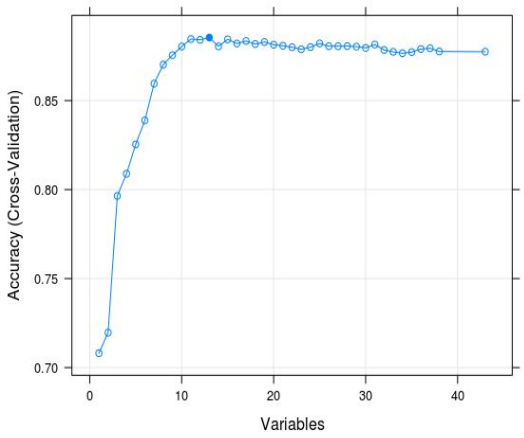


Figure 21: Number of selected features using Random forests and Cross-validation.

Figure 15 presents the selected features. The variables with symbol * are those that are highly correlated as shown in the correlation matrix in Figure 16. The accuracy of the model range

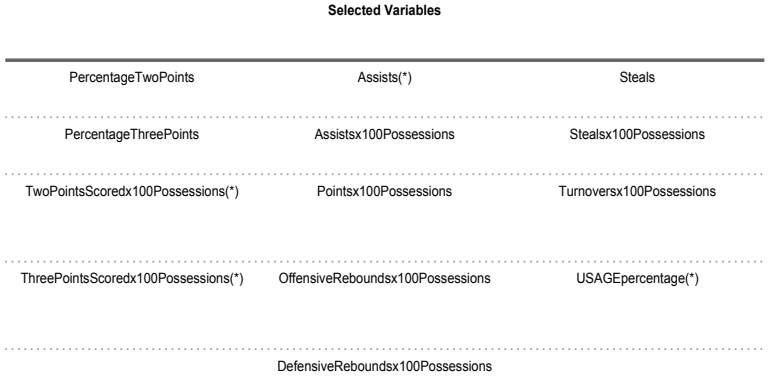


Figure 22: Selected features using Random forests and Cross-validation.

between $(0.854, 0.874)$. Table 1 shows the sensitivity and specificity and the positive and negative predictive values.

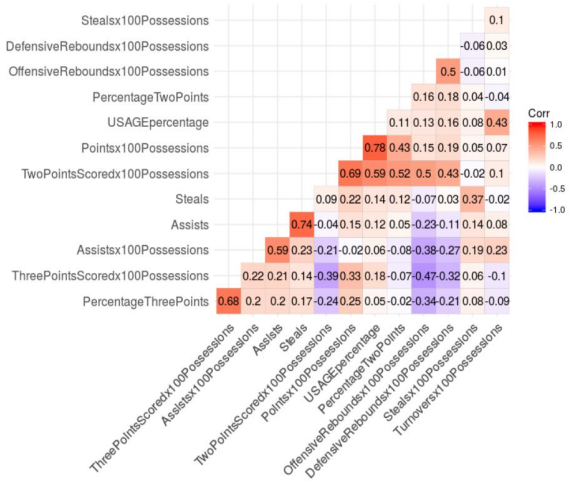


Figure 23: Correlation matrix of selected features.

Sensitivity	0.85
Specificity	0.88
Pos Pred Val	0.86
Neg Pred Val	0.87

Table 4: Predictive performance measures.

Recommendations and further directions

Most of the patterns found are well-known by specialists in the field of sports analytics in Basketball, but mainly in NBA where the is only one competition and the data has been collected during many years. We provided some statistical support to the data provided by Xpheres, with special focus on the comparison of several basketball leagues. It is important to mention that metrics are subjective, but they are very useful as it is not easy to find a overall global measure.

We recommend to include in the database some variables such as “experience” i.e. the number of professional years or other variables to account for the team and coach effects.

The proposed statistical analysis allows for a further validation step in the complete database. However, there are still many open questions.

Bibliography

[1] S.M. Berry, C.S. Reese and P.D. Larkey. "Bridging different eras in Sports", *Journal of the American Statistical Association.*Vol. 94, No 447 (1999), pp. 661–676.

[2] S. Bruce. "A Scalable Framework for NBA Player and Team Comparisons Using Player Tracking Data" (2016) https://arxiv.org/pdf/1511.04351.pdf

[3] D. Coates and B. Oguntimein. *"The Length and Success of NBA Careers: Does College Production Predict Professional Outcomes"*, Working Papers Series, Paper No. 08–06. International Association of Sports Economists (2008).

[4] P. Fearnhead. and B.M. Taylor. *"On estimating the Ability of NBA players"*. Journal of Quantitative Analysis in Sports. Vol. 7, Issue 3, Article 11, (2011).

[5] J. Kubatko, D. Oliver, K. Pelton and D.T. Rosenbaum. "A starting point for analyzing Basketball statistics", *Journal of Quantitative Analysis in Sports.*Vol. 3, Issue 3, Art 1 (2007).

[6] G.L. Page, B.J. Barney, and A.T. McGuire. "Effect of position, usage rate, and per game minutes played on NBA player production curves". *Journal of Quantitative Analysis in Sports*, Vol. 9, issue 4, Dec (2013).

[7] G.L. Page and F. Quintana. "Predictions based on the Clustering of Heterogeneous Functions via Shape and Subject-specific covariates". *Bayesian Analysis*, 10, number 2, pp. 379–410.

[8] S. Shea. Basketball Analytics: spatial tracking (2014). http://www.basketballanalyticsbook.com.

[9] R.P. Schumaker, O.K. Solieman and H. Chen. Sports Data Mining. Integrated Series in Information Systems 26 (2010). Springer.

List of participants

- Aleksandra Stojanova, Goce Delcev University of Stip (Macedonia)
- Ali Ramezani, BCAM - Basque Center for Applied Mathematics (Spain)
- Amaia Abanda Elustondo, BCAM - Basque Center for Applied Mathematics (Spain)
- Amaia Iparragirre Letamendia, BCAM & UPV/EHU (Spain)
- Antonio Zarauz Moreno, University of Almeria (Spain)
- Argyrios Petras, BCAM - Basque Center for Applied Mathematics (Spain)
- Bruno Flores Barrio, UR - University of La Rioja (Spain)
- Carlos Gorria Corres, UPV/EHU - University of the Basque Country (Spain)
- Christian Carballo Lozano, DeustoTech - Deusto Institute of Technology (Spain)
- Dae-Jin Lee, BCAM - Basque Center for Applied Mathematics (Spain)
- Dusan Bikov, Goce Delcev University of Stip (Macedonia)
- Edurne Iriondo Aznárez, University of Zaragoza (Spain)
- Elene Antón Balerdi, UPV/EHU - University of the Basque Country (Spain)
- Ferran Brosa Planella, University of Oxford (United Kingdom)
- Garritt Leland Page, BCAM & Brigham Young University (USA)
- Gorka Kobeaga, BCAM - Basque Center for Applied Mathematics (Spain)
- Gorka Labata Lezaun, University of Zaragoza (Spain)
- Iñigo Bidaguren, BCAM & UPV/EHU (Spain)
- Javier del Ser, BCAM & Tecnalia (Spain)
- Jesus Israel Epequin Chavez, IMJ - Institut de Mathématiques de Jussieu (France)
- Josu Doncel, UPV/EHU - University of the Basque Country (Spain)
- Laura Saavedra, BCAM & UPM (Spain)

- Manuel Higueras Hernáez, BCAM - Basque Center for Applied Mathematics (Spain)
- Mikel Lezaun Iturralde, UPV/EHU - University of the Basque Country (Spain)
- Mariam Kamal, UPV/EHU - University of the Basque Country (Spain)
- Mirjana Kocaleva, Goce Delcev University of Stip (Macedonia)
- Mostafa Shahriari, BCAM - Basque Center for Applied Mathematics (Spain)
- Nabil Fadai, University of Oxford (United Kingdom)
- Quan Wu, UPV/EHU - University of the Basque Country (Spain)
- Roi Naveiro Flores, ICMAT - Institute of Mathematical Sciences (Spain)
- Silvia García de Garayo Díaz, UPV/EHU - University of the Basque Country (Spain)
- Simón Rodríguez Santana, ICMAT - Institute of Mathematical Sciences (Spain)
- Thimjo Koca, Autonomous University of Barcelona (Spain)
- Thomas Ashley, IMUS - Institute of Mathematics of the University of Seville (Spain)
- Todor Balabanov, Bulgarian Academy of Sciences (Spain)

Acknowledgements

BCAM wishes to thank to the company speakers, the academic coordinators and the researchers of each working team for their invaluable contributions to the scientific success of the 131 European Study Group with Industry.

We also want to express our gratitude to BEAZ (public company of the Provincial Council of Bizkaia), UPV/EHU (University of the Basque Country) and MATH-IN (Spanish Network for Mathematics and Industry), as well as the institutions that financially supported the event: Basque Government, Severo Ochoa Excellence Accreditation and MI-NET Cost Action.