Data Envelopment Analysis of systems with multiple modes of functioning

S. Lozano and G. Villa*

Escuela Superior de Ingenieros, University of Seville

Camino de los Descubrimientos, s/n, 41092 Seville, Spain

* Corresponding author:

E-mail: gvilla@us.es

Phone: +34-954487207

This is a pre-copyedited, author-produced PDF of an article accepted for publication in Annals of Operations Research following peer review. The version of record G Villa, S Lozano (2024) Data envelopment analysis of systems with multiple modes of functioning, Annals of Operations Research, Volume 278, Issue 1-2, Pages 17 – 41, is available online at https://link.springer.com/article/10.1007/s10479-017-2733-7

Data Envelopment Analysis of systems with multiple modes of functioning

Abstract

Many systems can operate in different modes of functioning. Conventional Data Envelopment Analysis (DEA) would ignore that fact and consider instead that the system is a black box, paying attention just to the overall input consumption and output production. In this paper a more fine-grained approach is proposed consisting of explicitly modelling the different modes of functioning as specific processes and using the observed data on the input consumption and output production in each of the modes of functioning to infer the corresponding mode-specific technology. The system technology results from composing these mode-specific technologies according to the corresponding time allocations. The proposed approach allows computing efficient operating points for every mode of functioning, looking for improvements in the overall system performance. Two efficiency assessment DEA models are presented depending on whether the observed time allocation is maintained or the model is free to modify it. An application of the proposed approach to assessing the efficiency of NFL teams, operating in defence and offence modes in a given game, is presented.

Keywords: efficiency assessment; multiple modes of functioning; DEA; mode-specific technology; time allocative efficiency; NFL

Data Envelopment Analysis of systems with multiple modes of functioning

1. Introduction

Data Envelopment Analysis (DEA) is a non-parametric technique for evaluating the relative efficiency of homogeneous units commonly termed Decision Making Units (DMUs) (see, e.g., Cooper et al. 2000). Conventional DEA considers that the system under study is a black box whose input consumption and output production is, however, known. When the internal structure of the DMUs is known, a more fine-grained analysis is possible. This is what happens, for example, when the system consists on different processes, each one with its own inputs and outputs, and with intermediate products between the processes. For those systems a number of Network DEA models have been developed (e.g. Färe and Grosskopf 2000, Kao and Hwang 2008, Chen et al. 2009, Tone and Tsutsui 2009, Fukuyama and Weber 2010, Lozano 2011, 2015, 2016, Mirdehghan and Fukuyama 2016, etc). A review of Network DEA approaches was carried out in Kao (2014).

But Network DEA is not the only type of system with an internal structure. Thus, Castelli et al. (2010) also identify two other types: shared flow models and multilevel models. Shared flow models occur when some inputs or outputs are shared by different processes (e.g. Cook et al. 2000, Chen et al. 2010, Amirteimoori et al. 2016, Wu et al. 2016) while multilevel models (e.g. Cook et al. 1998) are considered when DMUs exhibits activities that cannot be associated to any of its processes.

In this paper a new type of internal structure DEA model is presented. It deals with the case in which the DMUs have multiple modes of functioning and operate a certain fraction of the time in each of these modes. Each mode of functioning (MF) can be considered as a process which consumes inputs and produces outputs. The overall input consumption and output production of the system is the aggregation of the inputs consumed and the outputs produced in all the different MFs used. The peculiarity is that the processes run on a time-sharing basis. Therefore, the performance of the whole system will be determined not only by the efficiency of the different MF but also by the amount of time that the system allocates to each MF. Consider for example the case of a reconfigurable manufacturing system, which can be set up to produce different part families using different tool types and fixtures. When the system is producing a part family, the system functions differently from when it produces

another part family. It can even use different types of input in each MF. Moreover, several MFs cannot run at the same time, i.e. when the system is dedicated to part family A, it cannot produce part family B. Another example would be the performance of an academic, who divides his/her time among different activities such as teaching, research and others. Each of these activities corresponds to a different MF, with its own inputs and outputs. Another application would be traffic regulation at intersections or in reversible lanes. Another application can be a toll road or bridge with the MFs corresponding to the number of toll booths open. This would be similar to considering the number of cash registers open in a supermarket as different MFs, also, the number of vehicles (and resulting headway) assigned to a route in an urban transit system. Following the dynamic transportation demand, different MFs can be used throughout a day.

The conventional DEA approach would ignore the existence of multiple MFs (disregarding their corresponding allocated time) and consider just the aggregate input consumption and output production. Our aim is just the opposite, i.e. to model the multiple MF (MMF) explicitly as specific processes, benchmarking the different processes using observed data about their inputs consumption and output production. Note that this is different from a parallel-process Network DEA approach (e.g. Kao 2009) due to the lack of simultaneity in the running of the different MFs, i.e. instead of operating in parallel the MFs operate using a time-sharing mechanism.

The structure of the paper is the following. In Section 2 the required notation is introduced and the mode-specific and overall MMF technologies are defined. In Section 3 the proposed MMF DEA models are presented. Section 4 presents a simple illustration of the proposed approach while Section 5 presents and discusses a real-world application to assess the efficiency of NFL teams with two MFs (defence and offence). Finally, the last section summarizes and concludes.

2. Production possibility set of MMF systems

Let us consider a certain physical device/system that can operate with M different modes of functioning. There is a set D of past observations (i.e. DMUs) so that each DMU j consists of the amount of inputs consumed $x_j^m = \left(x_{ij}^m\right)$, the amount of outputs produced $y_j^m = \left(y_{kj}^m\right)$ and the fraction of time t_j^m corresponding to each MF m. The usual notation of

indexes i and k is used above for the inputs and outputs. With no loss of generality, it is assumed that all MFs consume the same inputs $i \in I$ and produce the same outputs $k \in O$.

Figure 1b shows a graphical representation of a DMU j, with each MF represented as a box labelled "MF_#". Compare this scheme with that of Figure 1a, which corresponds to a conventional DEA approach (termed elementary DMU in Castelli et al. 2010) which would consider that the system is a black box, ignoring its MMF character. Such elementary DMU j would consume all the inputs of the different MF $x_{ij} = \sum_{m=1}^{M} x_{ij}^{m}$ and produce all the outputs of the different MF $y_{kj} = \sum_{m=1}^{M} y_{kj}^{m}$. For these elementary DMUs to be comparable these total inputs and outputs would have been the result of the operation of the system for a given time span $T \geq \sum_{m=1}^{M} t_j^m \quad \forall j$. Normally T is one week, one month, one year, etc. It is assumed that each DMU j has been idle (i.e. in no MF) for the corresponding time difference $T - \sum_{m=1}^{M} t_j^m$.

========== Figure 1 (about here) ==========

A Production Possibility Set (PPS) for each MF m can be determined from the available observations. This mode-specific PPS T^m can also be designated as the mode-specific technology of MF m. Consider the following axioms:

A.0. Null functioning:

$$\left(x^m = 0, y^m = 0, t^m = 0\right) \in T^m$$

A.1. Envelopment:

$$\left(x_j^m, y_j^m, t_j^m\right) \in T^m \quad \forall j \in D$$

A.2. Free disposability of inputs and outputs:

$$\left(x^m, y^m, t^m\right) \in T^m \wedge t^m > 0 \Rightarrow \left(\hat{x}^m, \hat{y}^m, t^m\right) \in T^m \quad \forall \hat{x}^m \geq x^m, \hat{y}^m \leq y^m$$

A.2'. Free disposability of functioning time:

$$\left(x^m, y^m, t^m\right) \in T^m \wedge t^m > 0 \Rightarrow \left(x^m, y^m, \hat{t}^m\right) \in T^m \quad \forall \hat{t}^m \geq t^m$$

A.3. Convexity:

$$\left.\begin{array}{l}\left(x^m, y^m, t^m\right) \in T^m \\ \left(\hat{x}^m, \hat{y}^m, \hat{t}^m\right) \in T^m\end{array}\right\} \Rightarrow \left(\alpha x^m + (1-\alpha)\hat{x}^m, \alpha y^m + (1-\alpha)\hat{y}^m, \alpha t^m + (1-\alpha)\hat{t}^m\right) \in T^m \ \forall 0 \leq \alpha \leq 1$$

A.4. Time scalability:

$$\left(x^m, y^m, t^m\right) \in T^m \wedge t^m > 0 \Rightarrow \left(\beta \frac{x^m}{t^m}, \beta \frac{y^m}{t^m}, \beta\right) \in T^m \quad \forall \beta \geq 0$$

The interpretation of the above axioms is the following. A.0 indicates that it is feasible for the system not to function in MMF m, of course without consuming any inputs or producing any outputs. A.1 indicates that the observed operation points of MMF m are feasible. A.2 indicates that, given a certain feasible operation time, it is possible, within the same functioning time, to waste inputs and outputs. A.2' indicates that, given a certain feasible operation time, it is possible to spend more time just to consume the same amount of inputs and produce the same amount of outputs. A.3 states that given two feasible operation points of MMF m, any convex linear combination of them, i.e. a mixture of both operating points, is also feasible. A.4 implies that, given a certain feasible operation time, any operation point of MMF m with the same rate of input of input consumption and output production is feasible.

Applying the Minimum Extrapolation Principle the following mode-specific PPS is obtained

$$T^m = \left\{(0,0,0)\right\} \cup \left\{ \begin{array}{l} \left(x^m, y^m, t^m\right): \\ \exists \lambda_j^m \geq 0 \ \forall j \in D \quad x^m \geq \sum_j \lambda_j^m \frac{x_j^m}{t_j^m} \quad y^m \leq \sum_j \lambda_j^m \frac{y_j^m}{t_j^m} \quad t^m \geq \sum_j \lambda_j^m > 0 \end{array} \right\} \tag{1}$$

Defining $\lambda_j^m = t^m \cdot \hat{\lambda}_j^m \quad \forall j \in D$, the above mode-specific PPS can be rewritten as

$$T^m = \left\{(0,0,0)\right\} \cup \left\{ \begin{array}{l} \left(x^m, y^m, t^m\right): \\ \exists \hat{\lambda}_j^m \geq 0 \ \forall j \in D \quad x^m \geq t^m \cdot \sum_j \hat{\lambda}_j^m \frac{x_j^m}{t_j^m} \quad y^m \leq t^m \cdot \sum_j \hat{\lambda}_j^m \frac{y_j^m}{t_j^m} \quad 1 \geq \sum_j \hat{\lambda}_j^m > 0 \end{array} \right\} \tag{1'}$$

The variable λ_j^m represents the length of time that the system operates as in MF m of each DMU j while $\hat{\lambda}_j^m$ expresses that length of time as a fraction of t^m. Also, the ratios $\frac{x_j^m}{t_j^m}$ and $\frac{y_j^m}{t_j^m}$ represent the input and output rates, respectively, of MF m of each observed DMU j.

The interpretation of the last inequality in (1') is that although A.4 implies Constant Returns to Scale with respect to functioning time (CRSwrtFT) the free disposability of the functioning time (i.e. time can be wasted) means that it is feasible to spend more time than the one that results from linearly combining the observed DMUs. However, although wasting time is feasible, it is not efficient. In other words, when looking for efficient operating points of MF m only the equality $\sum_j \hat{\lambda}_j^m = 1$ will do. And this is consistent with the fact that the above mode-specific PPS corresponds to Variable Returns to Scale with respect to inputs and outputs (VRSwrtIO).

To develop a mode-specific PPS corresponds to Constant Returns to Scale with respect to inputs and outputs (CRSwrtIO) we have to consider the following additional axiom:

A.5. Total input-output scalability:

$$\left(x^m, y^m, t^m\right) \in T^m \Rightarrow \left(\gamma x^m, \gamma y^m, t^m\right) \in T^m \quad \forall \gamma \geq 0$$

In that case then we arrive at the following CRSwrtIO mode-specific PPS

$$T^m_{CRSwrtIO} = \left\{\left(0,0,t^m\right) : t^m \geq 0\right\} \cup \left\{ \begin{array}{l} \left(x^m, y^m, t^m\right) : \exists \lambda^m_j \geq 0 \ \forall j \in D \quad \gamma^m_j \geq 0 \ \forall j \in D \\ x^m \geq \sum_j \lambda^m_j \gamma^m_j \frac{x^m_j}{t^m_j} \quad y^m \leq \sum_j \lambda^m_j \gamma^m_j \frac{y^m_j}{t^m_j} \quad t^m \geq \sum_j \lambda^m_j > 0 \end{array} \right\} \tag{1''}$$

This can be rewritten as

$$T^m_{CRSwrtIO} = \left\{\left(0,0,t^m\right) : t^m \geq 0\right\} \cup \left\{ \begin{array}{l} \left(x^m, y^m, t^m\right) : \exists \lambda^m_j \geq 0 \ \forall j \in D \quad \mu^m_j \geq 0 \ \forall j \in D \\ x^m \geq \sum_j \mu^m_j \frac{x^m_j}{t^m_j} \quad y^m \leq \sum_j \mu^m_j \frac{y^m_j}{t^m_j} \quad t^m \geq \sum_j \lambda^m_j > 0 \end{array} \right\} \tag{1'''}$$

Note that, in this CRSwrtIO case, the input-output components of the operation points of MMF are independent of the corresponding functioning time. This is because A.5 allows a trade-off between the functioning time and the rate of input consumption and output production. Thus, it is equivalent to function a certain time at certain input and output rates as to function half of that time at double rates or to function double that time at half the input and output rates. Note that A.5 is a very radical assumption, which allows attaining unbounded input and output rates for any functioning time. More reasonable seems to use the following alternative axiom

A.5'. Downward input-output scalability:

$$\left(x^m, y^m, t^m\right) \in T^m \Rightarrow \left(\gamma x^m, \gamma y^m, t^m\right) \in T^m \quad \forall\ 0 \leq \gamma \leq 1$$

In that case then we arrive at the following mode-specific PPS, which exhibits Non-Increasing Returns to Scale with respect to inputs and outputs (NIRSwrtIO)

$$T^m_{NIRSwrtIO} = \left\{\left(0,0,t^m\right): t^m \geq 0\right\} \cup \left\{\begin{array}{l}\left(x^m, y^m, t^m\right): \exists \lambda^m_j \geq 0 \ \forall j \in D \quad 0 \leq \gamma^m_j \leq 1 \ \forall j \in D \\ x^m \geq \sum_j \lambda^m_j \gamma^m_j \frac{x^m_j}{t^m_j} \quad y^m \leq \sum_j \lambda^m_j \gamma^m_j \frac{y^m_j}{t^m_j} \quad t^m \geq \sum_j \lambda^m_j > 0\end{array}\right\} \qquad (1^{iv})$$

This can be rewritten as

$$T^m_{NIRSwrtIO} = \left\{\left(0,0,t^m\right): t^m \geq 0\right\} \cup \left\{\begin{array}{l}\left(x^m, y^m, t^m\right): \exists \lambda^m_j \geq 0 \ \forall j \in D \quad \mu^m_j \geq 0 \ \forall j \in D \\ x^m \geq \sum_j \mu^m_j \frac{x^m_j}{t^m_j} \quad y^m \leq \sum_j \mu^m_j \frac{y^m_j}{t^m_j} \quad t^m \geq \sum_j \left(\lambda^m_j + \mu^m_j\right) > 0\end{array}\right\} \qquad (1^{v})$$

Note that A.5' is a relaxation of A.5 and implies that, given a feasible MMF m operation point, functioning at a fraction of the corresponding input-output rates (equivalent to slowing down the operation rate) is always feasible. Note also that the derivation of this last mode-specific PPS is similar to the way proposed in Kuosmanen (2005) for handling the weak disposability of undesirable outputs.

The efficient frontier of each mode-specific technology is formed by those feasible operation points that are non-dominated, i.e.

$$T^m_{eff} = \left\{(x,y,t) \in T^m : \neg\exists\left(\hat{x},\hat{y},\hat{t}\right) \in T^m \ \left(\hat{x},\hat{y},\hat{t}\right) \neq (x,y,t) \ \ \hat{x} \leq x \ \ \hat{y} \geq y \ \ \hat{t} \leq t\right\} \qquad (2)$$

The corresponding PPS of the MMF system (i.e. the overall MMF technology) is the composition/aggregation of the mode-specific technologies, i.e.

$$T^{MMF} = \left\{(x,y,t): \exists\left(x^m, y^m, t^m\right) \in T^m \ \forall m \quad x = \sum_m x^m \quad y = \sum_m y^m \quad t \geq \sum_m t^m\right\} \qquad (3)$$

Note that the above MMF technology takes into account that the system can spend time in an idle state in which no MF is running. In that state no input is consumed and no output is produced. The MMF efficient frontier is formed by those feasible operation points that are non-dominated, i.e.

$$T_{eff}^{MMF} = \left\{ (x,y,t) \in T^{MMF} : \neg\exists (\hat{x},\hat{y},\hat{t}) \in T^{MMF} \ \ (\hat{x},\hat{y},\hat{t}) \neq (x,y,t) \ \ \hat{x} \leq x \ \ \hat{y} \geq y \ \ \hat{t} \leq t \right\} \quad (4)$$

Note that the overall efficient operating points in T_{eff}^{MMF} never involve idleness, i.e.

$$(x,y,t) \in T_{eff}^{MMF} \Leftrightarrow \exists \left(x^m, y^m, t^m \right) \in T_{eff}^{m} \ \forall m \quad x = \sum_m x^m \quad y = \sum_m y^m \quad t = \sum_m t^m \quad (5)$$

3. MMF DEA efficiency assessment

Let 0 be the index of the DMU whose efficiency is to be assessed. Let us first formulate the corresponding DEA model if we consider it as an elementary DMU, i.e. ignoring its MMF structure (see Figure 1). CRSwrtFT and VRSwrtIO are assumed. The modifications required for NIRSwrtIO are straightforward. The modifications for the case of VRSwrtFT are not trivial and they are left as a topic for further research.

Because of its being a simple and flexible DEA metric, a Slacks-Based Inefficiency (SBI) measure will be used (Fukuyama and Weber 2009, 2010). Needless to say, the proposed approach can also be used with other DEA models, involving, for example, a radial, non-radial or Slacks-Based Efficiency (SBM), (Tone 2001, Tone and Tsutsui 2009) measure or a specific orientation (e.g. input, output or directional distance vector).

Let

$\hat{x}_i$ target amount of input i for DMU 0

$\hat{y}_k$ target amount of output k for DMU 0

$(\lambda_1, \lambda_2, ..., \lambda_n)$ intensity variables for linearly combining the observed DMUs

s_i^- slack for input i

s_k^+ slack for output k

g_i^x normalizing constant for slack of input i

g_k^y normalizing constant for slack of output k

Elementary DMU model (EM)

$$SBI_0^{EM} = \text{Max} \quad \left(\frac{1}{|I|} \sum_{i \in I} \frac{s_i^-}{g_i^x} + \frac{1}{|O|} \sum_{k \in O} \frac{s_k^+}{g_k^y} \right)$$

s.t.

$$\sum_j \lambda_j x_{ij} = \hat{x}_i \quad \forall i$$

$$\hat{x}_i = x_{i0} - s_i^- \quad \forall i \qquad (6)$$

$$\sum_j \lambda_j y_{kj} = \hat{y}_k \quad \forall k$$

$$\hat{y}_k = y_{k0} + s_k^+ \quad \forall k$$

$$\sum_j \lambda_j = 1$$

$$\lambda_j \geq 0 \quad \forall j \qquad s_i^-, s_k^+ \geq 0 \quad \forall i \, \forall k$$

Since we are considering VRSwrtIO the usual CRS DEA technology is considered, including the convexity constraint on the λ_j variables. The objective function corresponds to maximizing the sum of normalized input and output slacks (which reflect the existence of inefficiencies in the corresponding dimensions). Thus, DMU 0 would be considered efficient if $SBI_0^{EM} = 0$, while the larger the value of SBI_0^{EM} the more inefficient the DMU. A decomposition into input and output inefficiency measures can be made defining

$$SBI_0^{EM,x} = \frac{1}{|I|} \sum_{i \in I} \frac{s_i^-}{g_i^x}$$

$$SBI_0^{EM,y} = \frac{1}{|O|} \sum_{k \in O} \frac{s_k^+}{g_k^y} \qquad (7)$$

which leads to

$$SBI_0^{EM} = SBI_0^{EM,x} + SBI_0^{EM,y} \tag{8}$$

Instead of this conventional approach, the internal MMF structure of the DMUs can be taken into account and modelled. Two different MMF DEA models are proposed. In the first one, labelled MMF1, DMU 0 is projected onto the efficient frontier T_{eff}^{MMF} but maintaining the fraction of time that the system operates in each of the different MFs. In the second model, labelled MMF2, this requirement is relaxed and the model is free to search for an efficient operating point of the overall system with an improved allocation of time among the different MFs.

Let

$\hat{x}_i^m$ target amount of input i for MF m of DMU 0

$\hat{y}_k^m$ target amount of output k for MF m of DMU 0

$\left(\lambda_j^1, \lambda_j^2, ..., \lambda_j^M\right)$ intensity variables for linearly combining the MFs of the observed DMUs (j=1,2,..n)

Proposed MMF1 model

$$SBI_0^{MMF1} = \text{Max} \quad \left(\frac{1}{|I|} \sum_{i \in I} \frac{s_i^-}{g_i^x} + \frac{1}{|O|} \sum_{k \in O} \frac{s_k^+}{g_k^y} \right) \tag{9a}$$

s.t.

$$\sum_j \lambda_j^m \frac{x_{ij}^m}{t_j^m} = \hat{x}_i^m \quad \forall i \; \forall m \tag{9b}$$

$$\hat{x}_i = \sum_m \hat{x}_i^m = \sum_m x_{i0}^m - s_i^- \quad \forall i \in I \tag{9c}$$

$$\sum_j \lambda_j^m \frac{y_{kj}^m}{t_j^m} = \hat{y}_k^m \quad \forall k \, \forall m \tag{9d}$$

$$\hat{y}_k = \sum_m \hat{y}_k^m = \sum_m y_{k0}^m + s_k^+ \quad \forall k \tag{9e}$$

$$\sum_j \lambda_j^m = t_0^m \quad \forall m \tag{9f}$$

$$\lambda_j^m \geq 0 \quad \forall m \, \forall j \quad s_i^-, s_k^+ \geq 0 \quad \forall i \, \forall k \tag{9g}$$

The above model always has feasible solutions. Thus, for example, the solution $\lambda_j^m = 0 \; \forall j \neq 0 \, \forall m \;\; \lambda_0^m = t_0^m \;\; \forall m$, $\hat{x}_i^m = x_{i0}^m \;\; \forall i \;\; \forall m$, $\hat{y}_k^m = y_{k0}^m \;\; \forall k \;\; \forall m$, $s_i^- = 0 \;\; \forall i \;\; s_k^+ = 0 \;\; \forall k$ is feasible.

Note that, when computing the target operating point of MF m, the intensity variables λ_j^m represent the fraction of time that the system should operate as DMU j for that MF. Thus, the target results from replicating the operating points of the observed DMUs for that MF using the time allocation given by λ_j^m. The same as in a conventional DEA, only mode-specific efficient operating points can be used as benchmarks in the optimal linear combinations that define the target operating point of a certain MF. That is because the facets that form the mode-specific efficient frontier T_{eff}^m are defined by those efficient mode-specific DMUs. Mathematically, $\left(\lambda_j^m\right)^* > 0 \Rightarrow \left(x_j^m, y_j^m, t_j^m\right) \in T_{eff}^m$. Analogously, the target mode-specific operating point is efficient, i.e. $\left(\hat{x}^m, \hat{y}^m, t_0^m\right) \in T_{eff}^m \quad \forall m$.

Model MMF1 assesses the efficiency of a DMU by removing the inefficiencies in its different MFs but maintaining the observed time allocation. However, if the observed time allocation is not optimal there exists a time allocative inefficiency that can be removed letting the time spent in each MF as a decision variable to be determined by the DEA model. This leads to model MMF2 which is obtained from MMF1 defining

α^m fraction of time the target overall operating point should operate using MF m

and replacing (9f) by

$$\sum_j \lambda_j^m = \alpha^m \quad \forall m \tag{10a}$$

$$\sum_m \alpha^m = \sum_m t_0^m \tag{10b}$$

$$\alpha^m \geq 0 \quad \forall m \tag{10c}$$

The same as in model MMF1, the solution $\lambda_j^m = 0 \; \forall j \neq 0 \forall m \;\; \lambda_0^m = t_0^m \; \forall m$, $\hat{x}_i^m = x_{i0}^m \; \forall i \; \forall m$, $\hat{y}_k^m = y_{k0}^m \; \forall k \; \forall m$, $s_i^- = 0 \; \forall i \;\; s_k^+ = 0 \; \forall k$ is feasible in model MMF2. Actually, since MMF2 is a relaxation of model MMF1, every feasible solution of MMF1 is also feasible in MMF2. This means that $SBI_0^{MMF2} \geq SBI_0^{MMF1}$. In other words, MMF2 has more discriminant power than MMF1. Actually, the difference between the inefficiency scores computed by MMF2 and MMF1 is a measure of the time allocative inefficiency of DMU 0, i.e. the potential gain when the fraction of time during which the system operated in each MF is reallocated

$$SBI_0^{alloc} = SBI_0^{MMF2} - SBI_0^{MMF1} \tag{11}$$

leading to the following inefficiency decomposition

$$SBI_0^{MMF2} = SBI_0^{MMF1} + SBI_0^{alloc} \tag{12}$$

Note that (10b) implies that, although the target overall operating point can allocate time freely among the different MFs, the total operation time should be equal to that observed for DMU 0. Model MMF2 could even determine a target overall operating point specifying that the system should operate using a single MF. In any case, the target overall operating point is efficient, i.e. $\left(\hat{x}, \hat{y}, \sum_m t_0^m \right) \in T_{eff}^{MMF}$.

4. Illustration of proposed approach

In order to illustrate the proposed MMF models, let us consider a system with three MFs (labelled I, II and III) as shown in Figure 2. The system consumes two inputs and produces a single output. MFs I and III only consume input x_1 while MF II only consumes input x_2. Table 1 shows the data for four DMUs. The observed data include not only the amounts of input consumed and the amount of output produced in each MF but also the amount of time that each MF was used. Note that DMUs 2, 3 and 4 run all the time while DMU 1 was idle for 0.1 time units. Note also that DMU 3 did not use MF III.

========== Figure 2 (about here) ==========

========== Table 1 (about here) ==========

Assuming, for convenience in the numerical calculations, a slacks-normalizing vector $g = (g^x, g^y) = (1,1,1)$, the conventional DEA EM model for DMU 1 would be

$$SBI_1^{EM} = \text{Max} \ \frac{1}{2}(s_1^- + s_2^-) + s^+ \tag{13a}$$

s.t.

$$7\lambda_1 + 1\lambda_2 + 3\lambda_3 + 5\lambda_4 = \hat{x}_1 \tag{13b}$$

$$\hat{x}_1 = 7 - s_1^- \tag{13c}$$

$$3\lambda_1 + 2\lambda_2 + 5\lambda_3 + 2\lambda_4 = \hat{x}_2 \tag{13d}$$

$$\hat{x}_2 = 3 - s_2^- \tag{13e}$$

$$12\lambda_1 + 13\lambda_2 + 18\lambda_3 + 12\lambda_4 = \hat{y} \tag{13f}$$

$$\hat{y} = 12 + s^+ \tag{13g}$$

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1 \tag{13h}$$

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4, s_1^-, s_2^-, s^+ \geq 0 \tag{13i}$$

The optimal solution for the above model, as well as for the other three DMUs, is shown in Table 2. Note that only DMU 2 and DMU 3 are found to be efficient. DMU 4 is projected onto DMU 2 allowing a reduction of 4 units of input 1 and an increase of 1 unit of output. DMUs 1 are projected onto a convex combination of the two efficient DMUs 2 leading to larger input and output slacks which translate into a larger inefficiency score.

========== Table 2 (about here) ==========

As regards model MMF1 for DMU 1, this would be

$$SBI_1^{MMF1} = \text{Max} \ \frac{1}{2}(s_1^- + s_2^-) + s^+ \tag{14a}$$

s.t.

$$\frac{5}{0.6}\lambda_1^I + \frac{1}{0.3}\lambda_2^I + \frac{2}{0.5}\lambda_3^I + \frac{3}{0.2}\lambda_4^I = \hat{x}_1^I \tag{14b}$$

$$\frac{2}{0.1}\lambda_1^{III} + 0\lambda_2^{III} + \frac{1}{0.2}\lambda_3^{III} + \frac{2}{0.4}\lambda_4^{III} = \hat{x}_1^{III} \tag{14c}$$

$$\hat{x}_1^I + \hat{x}_1^{III} = 7 - s_1^- \tag{14d}$$

$$\frac{3}{0.2}\lambda_1^{II} + \frac{2}{0.7}\lambda_2^{II} + \frac{5}{0.3}\lambda_3^{II} + \frac{2}{0.4}\lambda_4^{II} = \hat{x}_2^{II} \tag{14e}$$

$$\hat{x}_2^{II} = 3 - s_2^- \tag{14f}$$

$$\frac{4}{0.6}\lambda_1^I + \frac{6}{0.3}\lambda_2^I + \frac{4}{0.5}\lambda_3^I + \frac{1}{0.2}\lambda_4^I = \hat{y}^I \tag{14g}$$

$$\frac{3}{0.2}\lambda_1^{II} + \frac{7}{0.7}\lambda_2^{II} + \frac{8}{0.3}\lambda_3^{II} + \frac{6}{0.4}\lambda_4^{II} = \hat{y}^{II} \tag{14h}$$

$$\frac{5}{0.1}\lambda_1^{III} + 0\lambda_2^{III} + \frac{6}{0.2}\lambda_3^{III} + \frac{5}{0.4}\lambda_4^{III} = \hat{y}^{III} \quad (14i)$$

$$\hat{y}^{I} + \hat{y}^{II} + \hat{y}^{III} = 12 + s^{+} \quad (14j)$$

$$\lambda_1^{I} + \lambda_2^{I} + \lambda_3^{I} + \lambda_4^{I} = 0.6 \quad (14k)$$

$$\lambda_1^{II} + \lambda_2^{II} + \lambda_3^{II} + \lambda_4^{II} = 0.2 \quad (14l)$$

$$\lambda_1^{III} + \lambda_2^{III} + \lambda_3^{III} + \lambda_4^{III} = 0.1 \quad (14m)$$

$$\lambda_1^{I},\lambda_2^{I},\lambda_3^{I},\lambda_4^{I},\lambda_1^{II},\lambda_2^{II},\lambda_3^{II},\lambda_4^{II},\lambda_1^{III},\lambda_2^{III},\lambda_3^{III},\lambda_4^{III},s_1^{-},s_2^{-},s^{+} \geq 0 \quad (14n)$$

As MF II does not consume input x_1 (see Table 1), the variable $\hat{x}_1^{II} = 0$ and it does not have to be included in the model. The same happens with variables $\hat{x}_2^{I}$ and $\hat{x}_2^{III}$. On the other hand, in (14d) variables $\hat{x}_1^{I}$ and $\hat{x}_1^{III}$ are summed in order to compute the total amount of resource x_1 consumed in the target solution provided by the model and hence the input slack with respect to total consumption observed. The same happens with output y in equation (14j), for all three MFs in this case.

Consider, for example, equation (14e) which defines the target of input x_2 for MF II in ($\hat{x}_2^{II}$) as the linear combination of inputs for all the observations. The ratios 3/0.2, 2/0.7, 5/0.3 and 2/0.4 represent the consumption of input x_2 per time unit for each DMU. The intensity variables $\lambda_1^{II},\lambda_2^{II},\lambda_3^{II},\lambda_4^{II}$ indicate what fraction of time the target should replicate each of the DMUs. The resulting linear combination of consumptions in MF II corresponds to the 0.2 duration of the use of MF II which is reflected in constraint (14l).

In order to formulate model MMF2, constraints (14k)-(14m) should be replaced by

$$\lambda_1^{I} + \lambda_2^{I} + \lambda_3^{I} + \lambda_4^{I} = \alpha^{I} \quad (15a)$$

$$\lambda_1^{II} + \lambda_2^{II} + \lambda_3^{II} + \lambda_4^{II} = \alpha^{II} \quad (15b)$$

$$\lambda_1^{III} + \lambda_2^{III} + \lambda_3^{III} + \lambda_4^{III} = \alpha^{III} \quad (15c)$$

$$\alpha^{I} + \alpha^{II} + \alpha^{III} = 0.9 \quad (15d)$$

$$\alpha^{I}, \alpha^{II}, \alpha^{III} \geq 0 \quad (15e)$$

These restrictions determine the value of the variables α^{I}, α^{II} and α^{III} which correspond to the fractions of time during which the MMF2 target of DMU 1 should operate in each mode, keeping the value of the observed idle time constant (0.1 in this case).

Tables 3 and 4 show the optimal values of the variables computed by MMF1 and MMF2. In both cases, DMU 2 is the only efficient DMU. The SBI scores of the inefficient DMUs are larger for MMF2 than for MMF1 leading to a different ranking from the one derived by EM. Thus, when the MMF structure is taken into account the inefficiency score of DMU 4 increases significantly while that of DMU 3 decreases.

Focusing on DMU 1, MMF1 provides $\left(\lambda_1^{I}, \lambda_2^{I}, \lambda_3^{I}, \lambda_4^{I}\right) = (0, 0.6, 0, 0)$, $\left(\lambda_1^{II}, \lambda_2^{II}, \lambda_3^{II}, \lambda_4^{II}\right) = (0, 0, 0, 0.2)$ and $\left(\lambda_1^{III}, \lambda_2^{III}, \lambda_3^{III}, \lambda_4^{III}\right) = (0.1, 0, 0, 0)$ (see Table 4), which means that, to be efficient, DMU 1 should operate in MF I as DMU 2 does, in MF II as DMU 4 does, and in MF III as DMU 1 does, maintaining the run times of these MFs. Doing this leads to a reduction of 3 and 2 units of inputs 1 and 2, respectively, as well as an increase of output y of 8 units (see Table 3). However, when MMF2 is applied, the only non-zero intensities are λ_1^{III}=0.17 and λ_3^{III}=0.73 (see Table 4), which means that, to be efficient, DMU 1 should operate the whole 0.9 time units exclusively in MF III and using an operating point that corresponds to replicating DMU 1 during 0.17 time units and DMU 3 during 0.73 time units. Doing this leads to a reduction of 3 units of input 2 and an increase of 18.33 of the output (see Table 3). The difference between the inefficiency scores computed by MMF2 and MMF1 is reported in the last column of Table 3 and corresponds to the time allocative inefficiency of the DMUs. In fact, DMU 1 is the one with the largest time allocative inefficiency.

========== Table 2 (about here) ==========

========== Table 3 (about here) ==========

========== Table 4 (about here) ==========

Note that model MMF2 computes an overall operating point, within the overall MMF technology, that optimizes the MF time allocation. Using that freedom, and based on the observed operation points of the different MFs, the model can determine that the optimal (in the SBI objective function sense) target operation point has a MF time allocation that may not make use of all MF and that, in general, can be quite different from the observed DMU being assessed. That should be interpreted as a MF time allocation inefficiency of the observed DMU, i.e. if the DMU had chosen the computed optimal MF time allocation and their corresponding target MF operation points, the sum of the input consumption reduction and the output production increases would be maximum.

Let us comment on two specific cases pointed out by one of the reviewers. Thus, the observed DMU 1 used all three MF (with a MF time allocation of 0.6, 0.2 and 0.1, respectively) producing a total of 12 units of output y and consuming a total of 7 and 3 units of inputs 1 and input 2, respectively. The application of MMF1, which respects the observed MF time allocation, represents an efficiency improvement as it allows achieving an output y=20 (i.e. 8 units more than DMU 1) with a consumption of 4 units of input 1 (3 units less than the observed DMU 1) and just 1 unit of input 2 (2 units less than DMU 1). Moreover, MMF2, which optimizes the MF time allocation, computes a feasible target operation point that uses just one MF (namely MF III) for the whole functioning time of 0.9, obtaining an output y=30.33 (18.33 more than the observed DMU 1) and consuming 7 units of input 1 (same as the observed DMU 1) and nothing of input 2. It is clear that this target overall operating point dominates both the original DMU 1 and the MMF1 target as it produces more output and consume less inputs.

As regards the observed DMU 2, it originally used MF I (functioning time $t_2^{I} = 0.3$) and MF II (functioning time $t_2^{II} = 0.7$). The MF I stint consumed $x_{12}^{I} = 1$ and produced $y_2^{I} = 6$ while MF II produced $y_2^{II} = 7$ and consumed $x_{22}^{II} = 2$. Overall the observed DMU 2

produced $y_2 = 13$ and consumed $x_{12} = 1$ and $x_{22} = 2$. Model MMF1, which respects the original MF time allocation, cannot improve the performance of the original DMU 2. Model MMF2 is not able either to find an overall operating point with more output and less input. Although model MMF2 computes an alternative optimum (that uses MF III instead of MF II but for the same length of time, producing the same amount of output and consuming the same amount of input), by looking at the zero optimal value of the SBI objective function it is clear that the original DMU 2 is efficient and that is why neither MMF1 nor MMF2 can find a feasible operating point that dominates it.

Therefore, the target overall operation point computed by MMF2 is generally quite different from the observed DMU being assessed. What matters, however, is that the optimal overall operation point can produce more outputs and consume less inputs than the observed DMU. If that happens, the inefficiency of the observed DMU is derived. If, on the contrary, no efficiency improvement can be attained (as it was the case with DMU 2 above), then, even if the target computed by MMF2 is different from the original DMU, the conclusion is that the original DMU is non-dominated and, therefore, efficient.

5. Application to efficiency assessment of NFL teams

In this section the proposed approach is applied to assess the efficiency of the 32 teams in the National Football League (NFL). The dataset used consists of the 16 weeks of the 2016 regular season and the 11 games played during the post-season (including the SuperBowl). Note that a football game is played between two teams of 11 players each and that teams may substitute any number of their players utilizing specialized offensive (offence unit), defensive (defence unit) and special team squads (the last is responsible for all kicking plays) (Quinn, 2012). Consequently, a football team can be seen as a system with two MFs: the offence and the defence. When the offence mode is used, the offence unit advances the ball down the field with the aim of scoring. When the team is operating in defence mode, the defence unit has the mission of preventing the offense unit of the rival team from scoring by tackling the ball carrier or by forcing turnovers. The possession of the ball determines the time that the team functions in the offence mode during the game. From a DEA perspective, the inputs used and the outputs produced in offence mode are different from the inputs used and outputs produced in defence modes. Figure 3 shows the MMF perspective of a football team playing a game, distinguishing its two MFs and the operating time in each MF.

========== Figure 3 (about here) ==========

The inputs and outputs considered have been classified into offence and defence category (see table 3). The inputs of the offence MF are non-discretionary. This is because their value is not determined by the team, but by the rival. For the same reason, the outputs of the defence MF are non-discretionary. In the offence mode the team tries to convert as many passes attempts; total first downs; kicking extra-points attempts; 2-point conversion attempts; net yards rushing and passing; and rival's penalty yards into points scored. On the contrary, in the defence mode, it tries to prevent rival's passes attempts; rival's total first downs; rival's kicking extrapoints attempts; rival's 2 points conversion attempts; rival's net yards rushing and passing; and penalty yards to be converted into points against. Table 4 shows the main statistics of the inputs and outputs considered for the 2016 regular season and post-season games.

========== Table 4 (about here) ==========

The notation to be used is the following. Let

I^{off} set of offensive inputs: passes attempts, total first downs, kicking extra-points attempts, 2-point conversion attempts, net yards rushing, net yards passing, and rival's penalty yards.

O^{def} set of defensive outputs: rival's passes attempts, rival's total first downs, rival's kicking extra-points attempts, rival's 2-point conversion attempts, rival's net yards rushing, rival's net yards passing, and own penalty yards.

x_{ij}^{off} amount of input $i \in I^{off}$ consumed by DMU j playing in offence MF.

y_{kj}^{def} amount of output $k \in O^{def}$ produced by DMU j playing in defence MF.

$TOFF_j$ no. of minutes played in offence mode (ball possession) for DMU j

$TDEF_j$ no. of minutes played in defence mode (rival's ball possession) for DMU j

sco_j number of points scored by DMU j

$rsco_j$ number of points against (rival's score) for DMU j

3.1. EM approach

The EM model does not distinguish whether the team plays in either offence or defence mode so that the inputs and outputs of the DMU are computed adding the inputs and outputs of both MFs. In order to formulate the model, let:

$(\lambda_1, \lambda_2, ..., \lambda_n)$ auxiliary variables used to compute linear combinations of the observed DMUs

s slack for points scored by DMU 0

t slack for points against DMU 0

EM Model

$$SBI_0^{EM} = \text{Max} \quad s + t \tag{16a}$$

s.t.

$$\sum_j \lambda_j x_{ij}^{off} \leq x_{i0}^{off} \quad \forall i \in I^{off} \tag{16b}$$

$$\sum_j \lambda_j rsco_j = rsco_0 - t \tag{16d}$$

$$\sum_j \lambda_j sco_j = sco_0 + s \tag{16e}$$

$$\sum_j \lambda_j y_{kj}^{def} \geq y_{k0}^{def} \quad \forall k \in O^{def} \tag{16f}$$

$$\sum_j \lambda_j = 1 \tag{16h}$$

$$\lambda_j \geq 0 \quad \forall j \qquad s, t \text{ integer} \tag{16i}$$

If the objective function (16a) is compared with that of (6) it can be noted that the two slacks s and t are normalized using $g^x = g^y = 1$. The interpretation of the corresponding SBI score is, thus, the sum of the additional points that could have been scored and the points against that could have been saved.

For the non-discretionary inputs and outputs no slack variables are considered and they do not appear in the objective function (see Banker and Morey, 1986). A final feature of this application is the integrality of the two slack variables, which guarantees the integrality of the corresponding variables points score and points against (see Lozano and Villa, 2006; Kuosmanen and Kazemi-Matin, 2009; Kazemi-Matin and Kuosmanen, 2009).

3.2. MMF approach

In the MMF approach two different set of auxiliary lambda variables need to be considered, one for each MF. These auxiliary variables represent the time that the DMU 0 being assessed should function as the corresponding MF of each observed DMU. Let:

$\left(\lambda_1^{off}, \lambda_2^{off}, ..., \lambda_n^{off}\right)$ auxiliary variables for offence MF

$\left(\lambda_1^{def}, \lambda_2^{def}, ..., \lambda_n^{def}\right)$ auxiliary variables for defence MF

MMF1 Model

$$SBI_0^{MMF1} = \text{Max} \quad s + t \tag{17a}$$

s.t.

$$\sum_j \lambda_j^{off} \frac{x_{ij}^{off}}{TOFF_j} \leq x_{i0}^{off} \qquad \forall i \in I^{off} \tag{17b}$$

$$\sum_j \lambda_j^{def} \frac{rsco_j}{TDEF_j} = rsco_0 - t \tag{17d}$$

$$\sum_j \lambda_j^{off} \frac{sco_j}{TOFF_j} = sco_0 + s \tag{17f}$$

$$\sum_j \lambda_j^{def} \frac{y_{kj}^{def}}{TDEF_j} \geq y_{k0}^{def} \quad \forall k \in O^{def} \tag{17h}$$

$$\sum_j \lambda_j^{off} = TOFF_0 \tag{17j}$$

$$\sum_j \lambda_j^{def} = TDEF_0 \tag{17k}$$

$$\lambda_j^{off}, \lambda_j^{def} \geq 0 \quad \forall j \quad s,t \quad \text{integer} \tag{17l}$$

Note that in the above MMF1 model the time allocation, i.e. the length of time that the team plays in each MF, is fixed and equal to the time allocation used by DMU 0.

Model MMF2 also uses a different set of auxiliary lambda variables for each MF and the interpretation of these variables is the same, i.e. the time that the team should function as the corresponding MF of each observed DMU. The only difference is that the MMF2 model computes the optimal MF time allocation, i.e. the time allocation that maximizes the improvements in points scored and in points against. Let

α^{off} time that DMU 0 should play in offence mode

α^{def} time that DMU 0 should play in defence mode

The MMF2 model (and corresponding optimal value SBI_0^{MMF2}) can be obtained from MMF1 model if equations (17j) and (17k) are replaced by:

$$\sum_j \lambda_j^{off} = \alpha^{off} \tag{18a}$$

$$\sum_j \lambda_j^{def} = \alpha^{def} \tag{18b}$$

$$\alpha^{off} + \alpha^{def} = TOFF_0 + TDEF_0 \tag{18c}$$

$$\alpha^{off}, \alpha^{def} \geq 0 \tag{18d}$$

3.3. Results obtained

The above models have been solved for 534 DMUs (corresponding to the 267 games played by local and visitor teams in the 2016 regular season and post-season). To compute normalized efficiency scores from the optimal solutions of the proposed EM and MMF models the following equations are used:

$$\theta_0 = 1 - \frac{s^* + t^*}{sco_0 + rsco_0} \tag{19a}$$

$$\theta_{0,off} = 1 - \frac{s^*}{sco_0 + rsco_0} \tag{19b}$$

$$\theta_{0,def} = 1 - \frac{t^*}{sco_0 + rsco_0} \tag{19c}$$

The overall and MF-specific efficiencies are related as per:

$$\theta_0 = \theta_{0,off} + \theta_{0,def} - 1 \tag{20}$$

The three panels in Figure 4 shows the box-plots of the overall, offensive and defensive efficiencies (19a)-(19c) provided by the EM, MMF1 and MMF2 models. It can be seen that, in most of the cases, the efficiency scores are high (the first quartile values are always higher than 0.75). EM has a low discriminant power, with as many as 76.97% of efficient DMUs (82.58% and 82.40% in the case of offensive and defensive efficiency, respectively). As expected MMF1 provides higher efficiency scores than MMF2, which is, thus, the model which has more discriminant power. Note also that for the two MMF models, although the average offensive efficiency is higher than the average defensive efficiency, the minimum value of defensive efficiency is higher than the minimum value of offensive efficiency (0.68 versus 0.42 for MMF1, 0.58 versus 0.33 for MMF2).

========== Figure 4 (about here) ==========

Figure 5 shows the value of the difference between the optimal time allocation α^{off} (i.e. the time in a game that the team should have played in an offensive mode) and the

observed possession of the ball (TOFF) versus the difference between the optimal time allocation α^{def} and the observed possession of the ball by the rival team (TDEF) for each of the 536 DMUs computed by MMF2 model. Note that all the observations are on the bisector line $y = -x$ because $\alpha^{def} - TDEF = -\left(\alpha^{def} - TOFF\right)$ as per equation (18c). The points with $\alpha^{def} - TDEF > 0$ correspond to cases in which teams should have played in the defence mode longer, while the points with $\alpha^{def} - TOFF > 0$ (which occurs more often) corresponds to the opposite, i.e. games in which the team should have played longer more in the offence mode. Not only it occurs more often that the team should have played longer in offence mode than the opposite, but also the difference between the optimal and the observed time allocations is larger in those occasions than in the opposite. Note also that this optimal time allocation can only be computed by the proposed MMF2 model.

========== Figure 5 (about here) ==========

Figure 6 shows the MMF2 efficiency scores of the two teams that played the 2016 Super Bowl. The Patriots won that game to the Falcons with a score of 34-28. The figure shows not only the efficiency score of that game but also of all the games that each team played along the regular season and post-season.

========== Figure 6 (about here) ==========

As regards the Super Bowl, it can be seen that the Patriots achieved a higher overall efficiency mainly due to a higher defensive efficiency. Note, however, that in the two prior games played by each of the two teams (against other rivals) it was the Falcons that achieve higher efficiency, and in particular, higher defensive efficiency. The fact that they failed to maintain that performance level against the Patriots in the Super Bowl may be the key to understand the final result. As regards the regular season, both teams had some ups and downs although the Falcons seem to have a more balanced efficiency record than the Patriots in both the offence and defence modes.

Figure 7 shows the MMF1 inefficiency score SBI_0^{MMF1} versus time allocative inefficiency SBI^{alloc} for the 534 DMUs. Note that, since in this application the variables points scored and against as well as the corresponding slack variables are integer, the SBI

inefficiency scores are also integer. That is why the points shown in Figure 7 resemble a grid. It can be seen that for most teams the SBI_0^{MMF1} score, which represents the number of additional points that the team should have scored plus the points against it could have saved, lies between 0 and 15. The vertical coordinate SBI^{alloc} represents the additional number of points scored and points against saved if in addition to the offence and defence modes being efficient, their corresponding time allocations were also optimized. It can be seen, that it is not uncommon for the MMF2 model to detect time allocative inefficiencies of 4 points, and sometimes even more. Particularly interesting are the points on the vertical axis. They correspond to DMUs that model MMF1 labeled as efficient for some of which model MMF2 was able to uncover some offence and/or defence inefficiencies.

========== Figure 7 (about here) ==========

4. Summary and conclusions

In this paper the efficiency assessment of systems which have different MFs is studied. Instead of ignoring this internal structure of the system, as the EM model does, a novel MMF approach, which explicitly models this situation is proposed. Using observed data a mode-specific technology can be inferred for each MF and the overall system technology results from composing these for any MF time allocation. Two variants, labelled MMF1 and MMF2, are considered, depending on whether the observed time allocation is maintained or is relaxed. The latter detects more inefficiency and therefore has more discriminant power. In any case, it can be argued that the proposed approach is more valid than the EM approach as it represents a perspective closer to the real functioning of the system and uses more fine-grained data, making a better use of the available information on how the real system works

The proposed approach is illustrated in detail by a simple 2-input/1-output example. In addition, the application of the proposed approach to the performance of NFL teams along the 2016 regular season and post-season is presented. The results confirm that the proposed MMF approach, especially the MMF2 model, has more discriminant power than conventional DEA. Moreover, the proposed MMF2 model not only provides efficiency scores and target input and output values but also computes an optimal time allocation between the different MF. The MMF perspective allows an enriched analysis of the performance of the system and its dependence on the performance of the different MF.

There are a number of topics that have not been addressed in this paper and that merit further research. Thus, the time scalability axiom implicitly implies CRSwrtFT. Other mode-specific technologies, exhibiting other returns to scale with respect to functioning time may be devised for MFs that have some type of warm up and/or shutdown periods. Also related to this, the proposed approach only takes into account the fraction of the total time that the MFs are operating, implicitly assuming that each MF runs once and for the given length of time. However, it can happen that the MFs are used in a dynamic fashion with the DMU switching between the different MFs as required. Including this, especially if it involves switching costs, is also a challenging question. Finally, this paper deals with efficiency assessment but the methodology can also be extended to planning the future operation of a system to attain certain output levels, using the observed data (at the MF level) to infer its overall PPS.

Acknowledgments

This research was carried out with the financial support of the Spanish Ministry of Science and the European Regional Development Fund (ERDF) grant DPI2013-41469-P. The authors would also like to thank the guest editors and the reviewers for their patience and constructive comments.

References

Amirteimoori, A., Kordrostami, S. and Azizi, H., "Additive models for network data envelopment analysis in the presence of shared resources", *Transportation Research Part D*, 48 (2016) 411-424.

Banker, R.D. and Morey, R. (1986). Efficiency analysis for exogenously fixed inputs and outputs, *Operations Research*, 34, 513-521

Barros, C.P. and Leach S., "Performance evaluation of the English Premier Football League with data envelopment analysis", *Applied Economics*, 38 (2006) 1449-1458.

Boscá, J.E., Liern, V., Martínez, A. and Sala R., "Increasing offensive or defensive efficiency? An analysis of Italian and Spanish football", *Omega,* 37 (2009) 63-78.

Castelli, L., Pesenti, R. and Ukovich, W., "A classification of DEA models when the internal structure of the Decision Making Units is considered", *Annals of Operations Research*, 173 (2010) 207-235.

Chen, Y., Cook, W.D., Li, N. and Zhu, J., "Additive efficiency decomposition in two-stage DEA", *European Journal of Operational Research*, 196 (2009) 1170-1176.

Chen, Y., Du, J., Sherman, H.D. and Zhu, J., "DEA model with shared resources and efficiency decomposition", *European Journal of Operational Research*, 207 (2010) 339-349.

Cook, W., Chai, D., Doyle, J. and Green, R., "Hierarchies and groups in DEA", *Journal of Productivity Analysis*, 10, 2 (1998) 177-198.

Cook, W., Hababou, M. and Tuenter, H.J.H., "Multicomponent Efficiency Measurement and Shared Inputs in Data Envelopment Analysis: An Application to Sales and Service Performance in Bank Branches", *Journal of Productivity Analysis*, 14 (2000) 209-224.

Cooper, W.W., Seiford L.M. and Tone, K., *Data Envelopment Analysis. A Comprehensive Text with Models, Applications, References and DEA-Solver Software*, Kluwer Academic Publishers, 2000.

Färe, R. and Grosskopf, S., "Network DEA", *Socio-Economic Planning Sciences*, 34 (2000) 35-49.

Fukuyama, H. and Weber, W.L., "A directional slacks-based measure of technical inefficiency", *Socio-Economic Planning Sciences,* 43 (2009) 274-287.

Fukuyama, H. and Weber, W.L., "A slacks-based inefficiency measure for a two-stage system with bad outputs", *Omega*, 38, 5 (2010) 398-409.

Kao, C., "Efficiency measurement for parallel production systems", *European Journal of Operational Research*, 196 (2009) 1107-1112.

Kao, C., "Network data envelopment analysis: A review", *European Journal of Operational Research*, 239, 1 (2014) 1-16.

Kao, C. and Hwang, S.N., "Efficiency decomposition in two-stage data envelopment analysis: An application to non-life insurance companies in Taiwan", *European Journal of Operational Research*, 185 (2008) 418-429.

Kazemi Matin, R. and Kuosmanen, T., "Theory of integer-valued data envelopment analysis under alternative returns to scale axioms", *Omega*, 37, 5 (2009) 988-995.

Kuosmanen, T., "Weak disposability in nonparametric production analysis with undesirable outputs", *American Journal of Agricultural Economics*, 87 (2005) 1077-1082

Kuosmanen, T. and Kazemi Matin, R., "Theory of integer-valued data envelopment analysis", *European Journal of Operational Research*, 192 (2009) 658-667.

Lozano, S. and Villa, G., "Data Envelopment Analysis of Integer-Valued Inputs and Outputs", *Computer and Operations Research*, 33, 10 (2006) 3004-3014.

Lozano, S., "Scale and cost efficiency analysis of networks of processes", *Expert Systems With Applications*, 38, 6 (2011) 6612-6617.

Lozano, S., "Alternative SBM Model for Network DEA", *Computers & Industrial Engineering*, 82 (2015) 33-40.

Lozano, S., "Slacks-based inefficiency approach for general networks with bad outputs. An application to the banking sector", *Omega*, 60 (2016) 73-84.

Mirdehghan, S.M. and Fukuyama, H., "Pareto–Koopmans efficiency and network DEA", *Omega*, 61 (2016) 78-88.

Quinn, K.G., “Field Position and Strategy in American Football”, in *The Oxford Handbook of Sports Economics: The Economics of Sports volume 1*, (L.H. Kahane and S. Shmanske, Eds.) 2012, DOI: 10.1093/oxfordhb/9780195387773.013.0011

Tone, K., “A slacks-based measure of efficiency in data envelopment analysis”, *European Journal of Operational Research*, 130 (2001) 498-509.

Tone, K. and Tsutsui, M., “Network DEA: A slacks-based measure approach”, *European Journal of Operational Research*, 197 (2009) 243-252.

Wu, J., Zhu, Q., Ji, X., Chu, J. and Liang, L., “Two-stage network processes with shared resources and resources recovered from undesirable outputs”, *European Journal of Operational Research*, 251 (2016) 182-197

List of table and figure captions

	MF I				MF II				MF III				TOTAL			
DMU	x_1^I	x_2^I	y^I	t^I	x_1^{II}	x_2^{II}	y^{II}	t^{II}	x_1^{III}	x_2^{III}	y^{III}	t^{III}	x_1	x_2	y	$t^I + t^{II} + t^{III}$
1	5	-	4	0.6	-	3	3	0.2	2	-	5	0.1	7	3	12	0.9
2	1	-	6	0.3	-	2	7	0.7	-	-	-	-	1	2	13	1.0
3	2	-	4	0.5	-	5	8	0.3	1	-	6	0.2	3	5	18	1.0
4	3	-	1	0.2	-	2	6	0.4	2	-	5	0.4	5	2	12	1.0

Table 1. Observed data (4 DMUs) for system with 3 MFs

DMU	λ_1	λ_2	λ_3	λ_4	$\hat{x}_1$	$\hat{x}_2$	$\hat{y}$	s_1^-	s_2^-	s^+	SBI^x	SBI^y	SBI
1	0	0.67	0.33	0	1.67	3	14.67	5.33	0	2.67	2.67	2.67	5.33
2	0	1	0	0	1	1	13	0	0	0	0	0	0
3	0	0	1	0	3	5	18	0	0	0	0	0	0
4	0	1	0	0	1	2	13	4	0	1	2	1	3

Table 2. EM solution: intensity variables, input and output targets, input and output slacks and inefficiency scores

	DMU	Targets									Slacks			Inefficiency scores		
		MF I			MF II			MF III								
MMF1		$\hat{x}_1^I$	$\hat{x}_2^I$	$\hat{y}^I$	$\hat{x}_1^{II}$	$\hat{x}_2^{II}$	$\hat{y}^{II}$	$\hat{x}_1^{III}$	$\hat{x}_2^{III}$	$\hat{y}^{III}$	s_1^-	s_2^-	s^+	SBI^x	SBI^y	SBI
	1	2	-	12	-	1	3	2	-	5	3	2	8	2.5	8	10.5
	2	1	-	6	-	2	7	-	-	-	-	-	-	0.0	0.0	0.0
	3	1.67	-	10	-	1.5	4.5	1.33	-	6.44	-	3.5	2.94	1.75	2.94	4.69
	4	0.67	-	4	-	2	6	4.33	-	15.11	-	-	13.11	0	13.11	13.11

	DMU	Targets									Slacks			Inefficiency scores			
		MF I			MF II			MF III									
MMF2		$\hat{x}_1^I$	$\hat{x}_2^I$	$\hat{y}^I$	$\hat{x}_1^{II}$	$\hat{x}_2^{II}$	$\hat{y}^{II}$	$\hat{x}_1^{III}$	$\hat{x}_2^{III}$	$\hat{y}^{III}$	s_1^-	s_2^-	s^+	SBI^x	SBI^y	SBI	SBI^{alloc}
	1	-	-	-	-	-	-	7	-	30.33	-	3	18.33	1.5	18.33	19.83	9.33
	2	-	-	-	-	2	7	1	-	6	-	-	-	0.0	0.0	0.0	0.0
	3	-	-	-	-	5	9	3	-	18	-	-	9	0.0	9	9	4.31
	4	-	-	-	-	-	-	5	-	30	-	2	18	1	18	19	5.89

Table 3. Targets, slacks and inefficiency scores computed by MMF1 and MMF2 models

	DMU	MF I					MF II					MF III				
		λ_1^I	λ_2^I	λ_3^I	λ_4^I	$\sum_j \lambda_j^I$	λ_1^{II}	λ_2^{II}	λ_3^{II}	λ_4^{II}	$\sum_j \lambda_j^{II}$	λ_1^{III}	λ_2^{III}	λ_3^{III}	λ_4^{III}	$\sum_j \lambda_j^{III}$
MMF1	1	-	0.6	-	-	0.6	-	-	-	0.2	0.2	0.1	-	-	-	0.1
	2	-	0.3	-	-	0.3	-	0.7	-	-	0.7	-	-	-	-	-
	3	-	0.5	-	-	0.5	-	-	-	0.3	0.3	0.02	-	0.18	-	0.2
	4	-	0.2	-	-	0.2	-	-	-	0.4	0.4	0.16	-	0.24	-	0.4

	DMU	MF I					MF II					MF III					$\alpha^I + \alpha^{II} + \alpha^{III}$
		λ_1^I	λ_2^I	λ_3^I	λ_4^I	α^I	λ_1^{II}	λ_2^{II}	λ_3^{II}	λ_4^{II}	α^{II}	λ_1^{III}	λ_2^{III}	λ_3^{III}	λ_4^{III}	α^{III}	
MMF2	1	-	-	-	-	-	-	-	-	-	-	0.17	-	0.73	-	0.9	0.9
	2	-	-	-	-	-	-	0.7	-	-	0.7	-	0.1	0.2	-	0.3	1.0
	3	-	-	-	-	-	-	-	0.26	0.14	0.4	-	-	0.6	-	0.6	1.0
	4	-	-	-	-	-	-	-	-	-	-	-	-	1.0	-	1	1.0

Table 4. Intensity variables and time allocation computed by MMF1 and MMF2 models

OFFENCE VARIABLES		DEFENCE VARIABLES	
INPUTS	**LABEL**	**OUTPUTS**	**LABEL**
Passes Attempts	**PA**	Rival´s Passes Attempts	**RPA**
Total First Downs	**TFD**	Rival´s Total First Downs	**RTFD**
Kicking Extra-points Attempts	**KEPA**	Rival´s Kicking Extra-points Attempts	**RKEPA**
2-Point Conversion Attempts	**2PCA**	Rival´s 2-Point Conversion Attempts	**R2PCA**
Field Goals Attempts	**FGA**	Rival´s Field Goals Attempts	**RFGA**
Net Yards Rushing	**NYR**	Rival´s Net Yards Rushing	**RNYR**
Net Yards Passing	**NYP**	Rival´s Net Yards Passing	**RNYP**
Rival´s Penalty Yards	**RPY**	Penalty Yards	**PY**
OUTPUT	**LABEL**	**INPUT**	**LABEL**
Score	**SCO**	Rival's Score	**RSCO**
TIME ALLOCATION	**LABEL**	**TIME ALLOCATION**	**LABEL**
Offence time	**TOFF**	Defence time	**TDEF**

Table 4. Variables used in NFL application

Variable	PA	TFD	KEPA	2PCA	FGA	NYR	NYP	PY	SC	TOFF
Mean	36.49	19.58	2.09	0.25	1.97	103.61	236.83	60.90	21.37	29.81
St. dev.	8.29	5.07	1.33	0.51	1.36	47.87	73.03	29.59	9.31	4.45
Max	63	37	6	3	6	249	498	200	48	44.20
Min	15	1	0	0	0	6	6	0	0	18.53
Variable	**RPA**	**RTFD**	**RKEPA**	**R2PCA**	**RFGA**	**RNYR**	**RNYP**	**RPY**	**RSC**	**TDEF**
Mean	35.03	21.36	2.58	0.18	1.96	114.40	247.31	54.35	24.36	30.66
St. dev.	7.88	7.91	1.37	0.48	1.25	63.53	70.92	26.33	9.25	4.23
Max	58	125	7	4	6	764	481	145	49	46.35
Min	18	8	0	0	0	14	63	5	0	19.72

Table 5. Summary statistics of variables

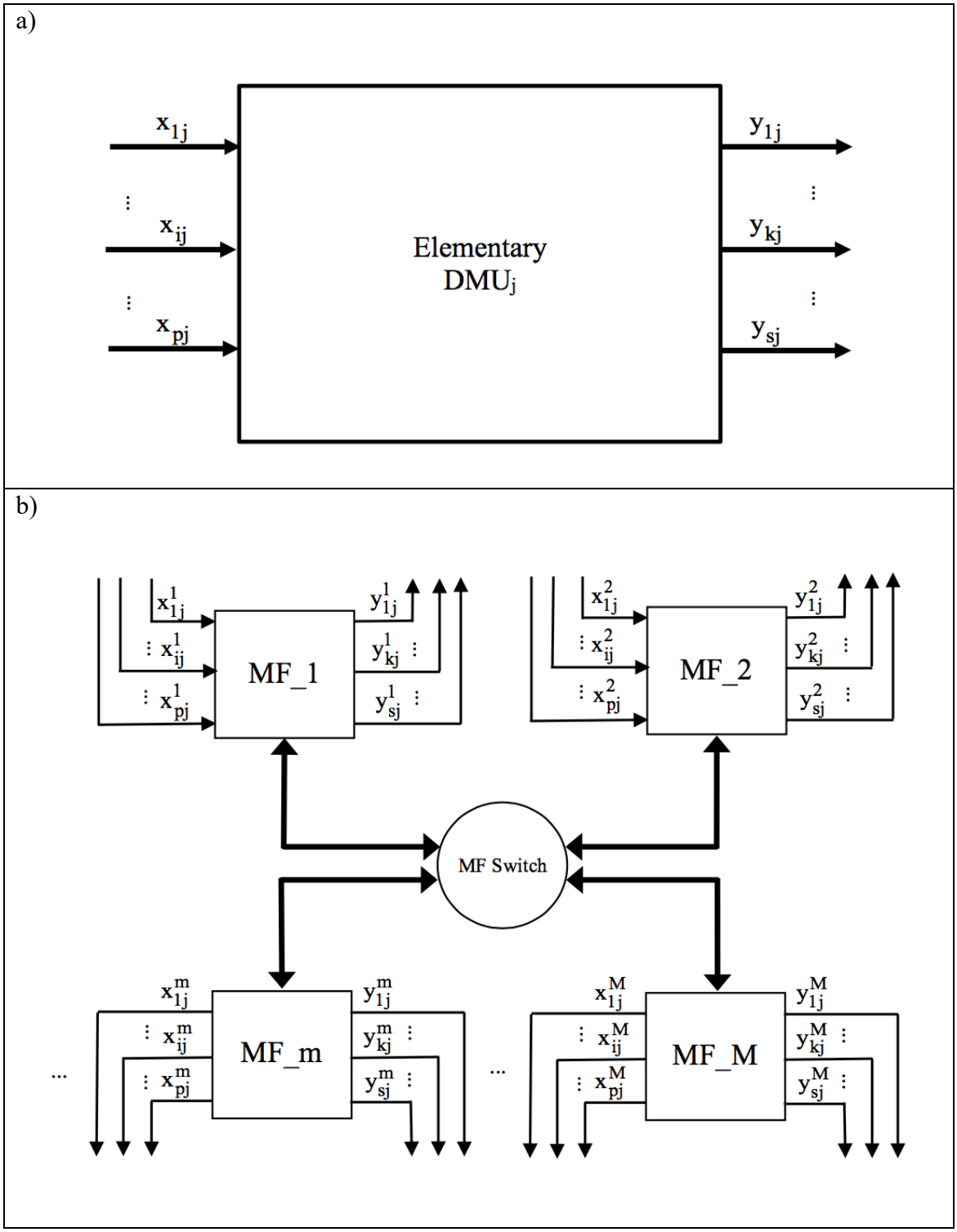

Figure 1. MMF system: a) EM perspective b) Proposed perspective

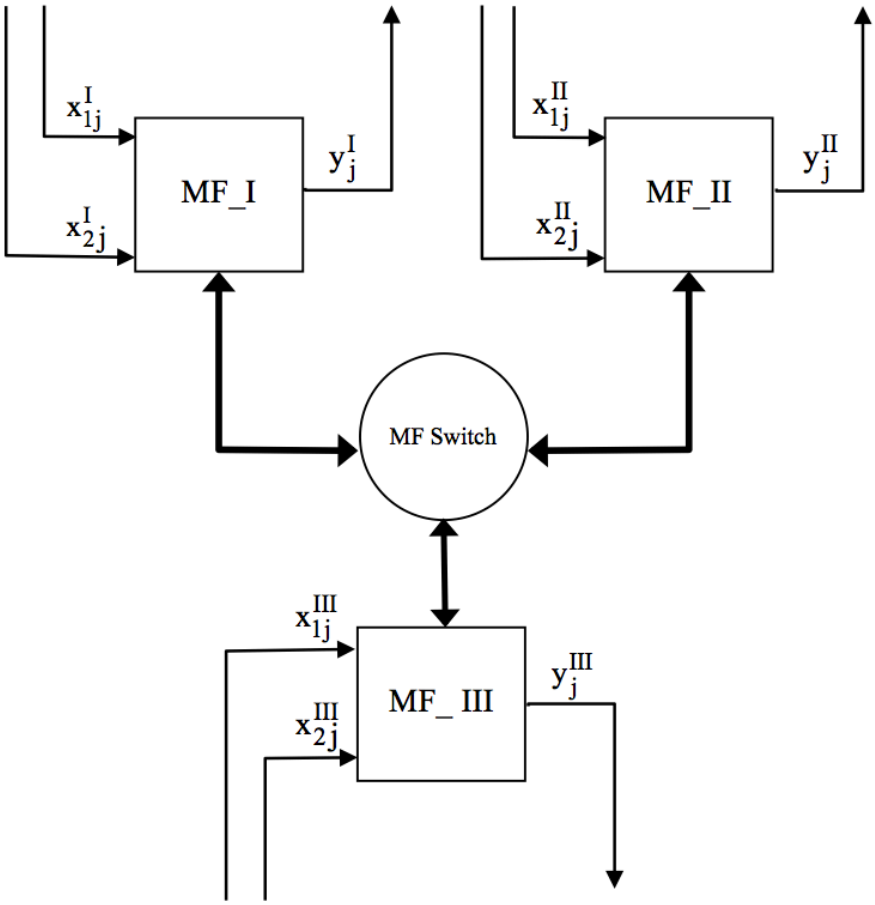

Figure 2. Illustration: system with 3 MFs, 2 inputs and a single output

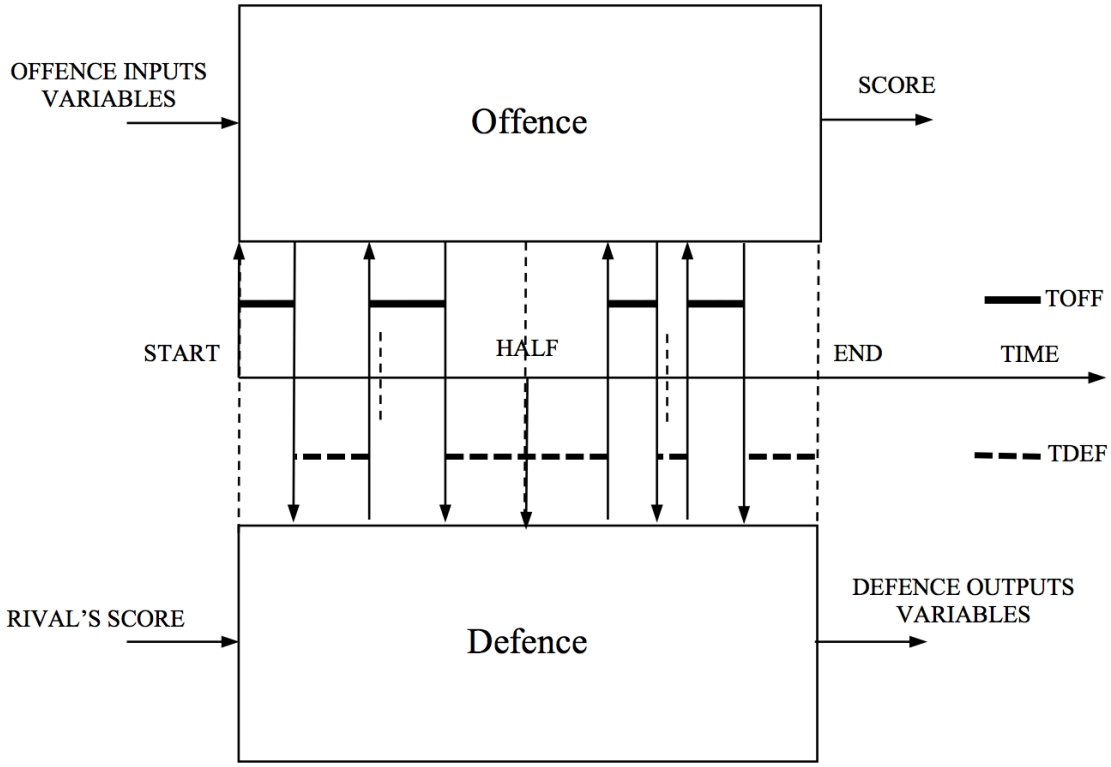

Figure 3. MMF perspective of NFL teams

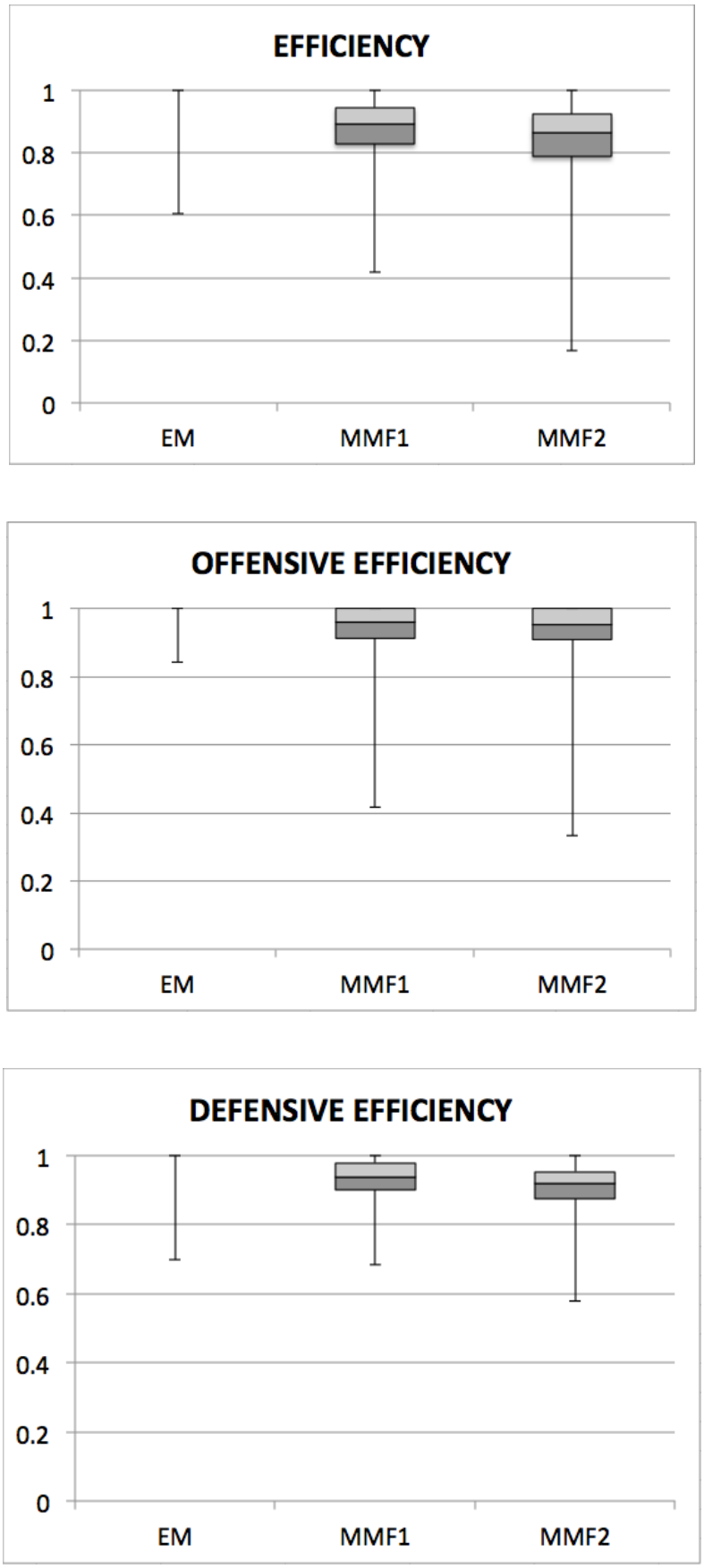

Figure 4. Box-plots of overall, offensive and defensive efficiency scores computed by EM, MMF1 and MMF2 models

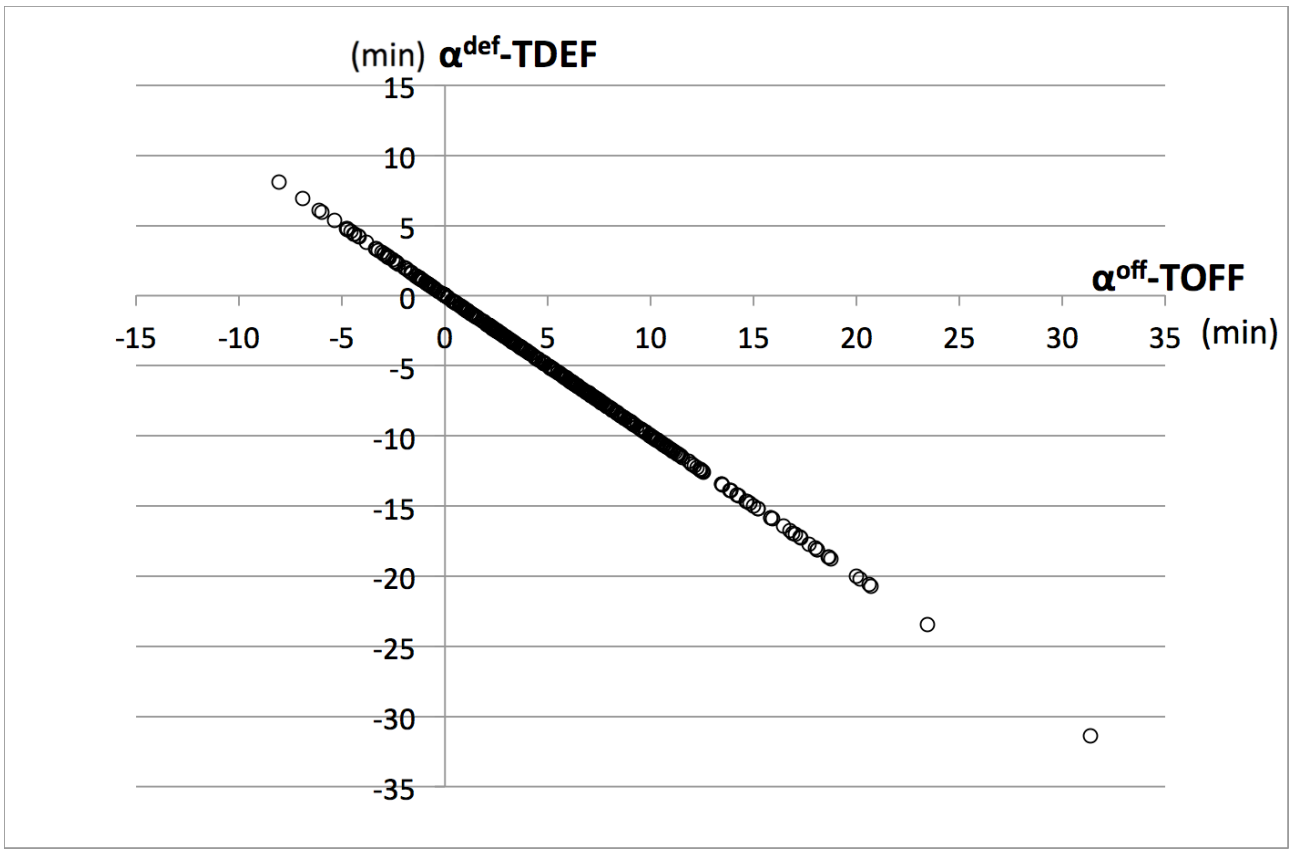

Figure 5. Difference between optimal and observed time allocations for the 534 DMUs

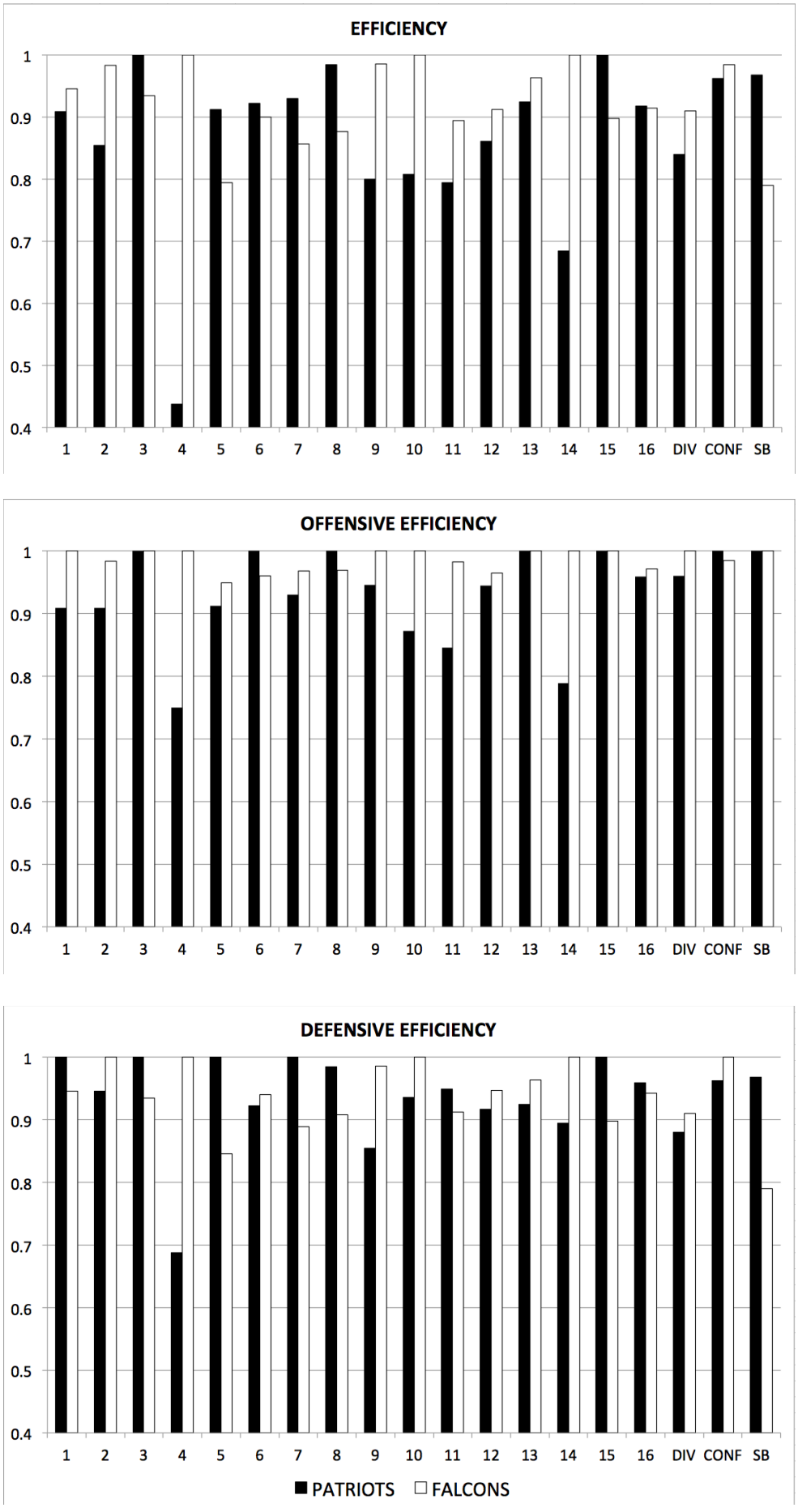

Figure 6. Overall, offensive and defensive efficiency scores computed by MMF2 model for the games played by the Patriots and by the Falcons during the 2016 Regular Season, Divisional Playoffs (DIV), Conference Championship (CONF) and Super Bowl (SB)

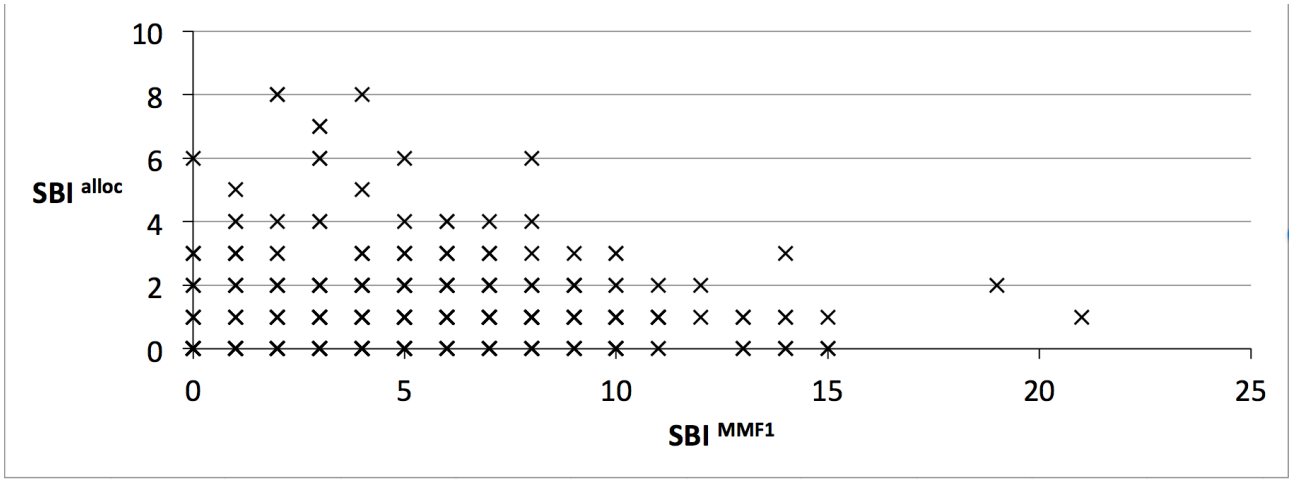

Figure 7. MMF1 inefficiency score SBI^{MMF1} versus time allocative inefficiency SBI^{alloc} for the 534 DMUs