Ann Oper Res (2010) 181: 287–302
DOI 10.1007/s10479-010-0743-9

The dynamics of performance space of Major League Baseball pitchers 1871–2006

Wen-Chih Chen · Andrew L. Johnson

Published online: 25 April 2010

Abstract A central decision-maker, the principal, employs performance evaluation criteria consistent with an organization's overall goal(s) to measure the effectiveness of the agents who execute decisions and implement strategies within a specified period. The dataset of performance criteria spanning the performance space will change over time to reflect the principal's strategic modifications. This paper applies a Data Envelopment Analysis (DEA) based approach to reveals the dynamics of the performance space of Major League Baseball (MLB) pitchers with minimum subjective judgment imposed on the data. The proposed approach is applied to data on MLB pitchers from 1871 to 2006. We conclude that many of the findings are consistent with the observations of baseball's experts. The approach also suggests new directions for investigating a large dataset to identify revealed preferences or strategies by using historical and modern observations.

Keywords Performance space · Dynamics · Data Envelopment Analysis

1 Introduction

In performance evaluation often multiple criteria should be considered. Typically these criteria are determined by a central decision-maker (in this paper called the principal). Multiple criteria may be used to reflect the multiple objectives of the principal as organizational leader or to reflect an acknowledgement of the many choices available to achieve an objective. The agents employed by the principal execute the appropriate strategies selected, and the recorded results of their performance over time are observed and collected.

W.-C. Chen
Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu, Taiwan
e-mail: wenchih@faculty.nctu.edu.tw

A.L. Johnson (✉)
Industrial and Systems Engineering Department, Texas A&M University, College Station, TX, USA
e-mail: ajohnson@tamu.edu

Technically speaking, the set of criteria that span the defined performance space are called performance dimensions. This set may change over time as the principal makes strategic adjustments to achieve objectives. These modifications may be motivated by the environment, regulations, operational changes, or simply a change in preference. The importance of the criteria not only changes, but also the set of criteria themselves may change.

The change in the performance criteria over time is called the dynamics of performance space. In this paper, "dynamics" means both a new dimension (criterion) added to the performance space as well as a potential reduction in dimension. We model the dynamics of the performance space using the DEA based approach developed in Pastor et al. (2002). This new interpretation of Pastor's model allows the inference of the strategic behaviors of the principals and the agents from the changes observed extending the original purpose to allow knowledge discover regarding the identification of strategic behavior. DEA considers multiple aspects of the performance simultaneously and aggregates different criteria values without *a priori* assumptions regarding weight assignments or functional form. DEA methods have been used to investigate various aspects of baseball performance, for example, Anderson and Sharp (1997), Lewis and Sexton (2004), and Hadley and Ruggiero (2006). However, our approach objectively reveals the dynamics of the performance space endogenously determining the appropriate criteria.

We apply this approach to Major League Baseball (MLB) pitchers' performance evaluations from 1871 to 2006 where team managers are the principles and pitchers are the agents. The results reveal how the relevant criteria for evaluating pitchers' performances change over the time period. We find that "games" and "innings pitched" are important in pitching performance evaluations over the entire dataset and "earned runs", "hits allowed" and "shutouts" are not significant in determining performance. Some performance measures such as "wins", "strikeouts", "complete game" and "saves" become key criteria in specific time periods. We conclude that our findings are highly consistent with baseball experts' observations, despite being developed independently and objectively. These results support the novel implementation of Pastor's statistical test for nested radial models to analyze and identify changing evaluation criteria over time.

The approach suggests that new directions can be taken to investigate a dataset to identify revealed preferences or strategies through the use of historical and modern observations. We note that the directions are not necessarily straightforward in conventional knowledge discovery processes. This approach can support prior hypotheses based upon intuition or experience, or can be utilized to discover unexpected knowledge. While the approach is applied to a sports league, we believe it can also be used wherever the performance of agents is guided by a principal's strategic behaviors to achieve an overall goal. In this sense, employees in a firm, branches of a chain store, or divisions within an academic or governmental entity could be investigated.

The remainder of this paper is organized as follows. Section 2 introduces DEA as a performance evaluation method. Section 3 addresses the criteria selection procedure. Sections 4 and 5 present an empirical study of MLB pitchers and the observations of performance criteria. Sections 6 and 7 investigate the sensitivity of the parameters selected and the effects of analyzing the American and National leagues separately. Section 8 concludes.

2 Data Envelopment Analysis (DEA)

Typically, performance evaluation is based upon multiple criteria. Consider a criteria set $C = I \cup O$ where I contains criteria to be minimized, i.e. smaller values are valued more,

and O contains criteria to be maximized. The notation, y_{jk} is the performance level for a particular observation k for $j \in O$ that positively influences the performance of the observation. Further, x_{ik} is the performance level for observation k for the criterion $i \in I$ that characterizes the opportunity and negative outcomes of the observation. One way to present the overall performance for record k is measured as the ratio of weighted sum of maximizing criteria to weighted sum of minimizing criteria. Determining the weights is essential but difficult. Despite many methods and studies on aggregating multiple criteria using weights into overall performance, the selection of the set weights often depends on personal experiences and subjective opinions.

DEA, popularized by Charnes et al. (1978) is a method to evaluate the relative performance by peer comparison. In particular, DEA considers multiple aspects of the performance simultaneously without *a priori* assumptions regarding weight assignments or functional form. Suppose K is the dataset with the performance of all agents. Banker et al. (1984) propose to aggregate overall performance in the following form:

$$\frac{\sum_{j\in O} u_j y_{jk} - u_\alpha}{\sum_{i\in I} v_i x_{ik}}. \tag{1}$$

u_j and v_i are the weights associated with criteria i and j respectively, and the scalar variable u_α insures that records are only compared to records of similar size. To evaluate the record k, the DEA problem can be formulated as:

$$\begin{aligned} \theta_k(C) = \max_{u,v}\ & \frac{\sum_{j\in O} u_j y_{jk} - u_\alpha}{\sum_{i\in I} v_i x_{ik}} \\ \text{s.t.}\ & \frac{\sum_{j\in O} u_j y_{jr} - u_\alpha}{\sum_{i\in I} v_i x_{ir}} \le 1, \quad \forall r \in K; \\ & u_j \ge 0, \qquad v_i \ge 0, \quad \forall i \in I, \forall j \in O; \\ & u_\alpha \text{ is free.} \end{aligned} \tag{2}$$

The programming problem (2) evaluates performance based on (1). Instead of selecting a common set of weights, (2) determines the set of weights for observed record k which is under evaluation, allowing k to achieve the highest possible performance score subject to a normalization constraint on the performance score of all observations. This programming problem must be solved once for each observation. The results are performance scores $\theta_k(C)$ for all records.

In most literature, DEA is used to provide relative efficiencies; we refer to DEA's ratio form and associated interpretation as a normalized overall performance score. Among various multi-criteria performance evaluation approaches, DEA uses a minimal set of assumptions concerning the form and weights and thus lends the model flexibility to quantify the performance of the observations characterized by the data over the criteria space. Simply stated, DEA offers a more objective assessment of performance relative to expert opinion or other common aggregation methods.

3 Determining the criteria for a performance model

Criteria selection is often a difficult process. Golany and Roll (1989) and Dyson et al. (2001) provide comprehensive discussions on the application of the DEA performance assessment

method and possible pitfalls. In practice, expert opinion often allows for some level of subjectivity to influence the criteria selection process. While it is usually expected that experts will agree on the significant indicators, differences of opinion invariably arise as more indicators are included, or as it becomes more difficult to finalize the total of indicators. As a nonparametric relative performance method DEA suffers from the curse of dimensionality. A parsimonious model is desirable since DEA's discriminating power decreases as the dimensions of the performance space increase (Simar and Wilson 2008). While some criteria that have an influence on performance score may be excluded from the model, the benefits in terms of discriminating power outweigh the slight information gain when including additional measures.

There are several different methods for selecting criteria in DEA performance evaluation model, for example, Casu et al. (2005), Cinca and Molinero (2004), Cook and Zhu (2007), Pastor et al. (2002) and Wanger and Shimshak (2007). We recognize that while no single model specification procedure will work for all cases, a method with minimum subjective judgments imposed on the data and based on a statistical test considering data randomness is preferred. It is often easy to identify a relatively large set of criteria that potentially could be important. One approach for model specification is to identify the large set and remove criteria that will not impact the overall performance. This is referred to as a model reduction or backward elimination procedure. Pastor et al. (2002) propose a model reduction method satisfying the requirements, which is also applied by Lovell and Pastor (1997) and Pastor et al. (2006) to studying financial institutions. Our approach extends Pastor et al. (2002) to analyze performance criteria changes over time and can contribute additional support to experts' opinions and assist with decisions regarding marginally important criteria and the scope of the criteria to be included.

Pastor et al. suggests using a backward elimination procedure to remove criteria that do not contribute significantly to the overall performance as measured by (2). The authors quantify a criterion's marginal impact on an observation as the change in overall performance when a particular criterion is included in the model compared to the performance level with the criterion excluded. If the difference for an observation is more than the pre-specified threshold, the specific observation is said to be *affected* by the existence of the criterion. A criterion does not have significant influence or is not important if only a few observations are affected, i.e., the proportion of affected observations is less than a predetermined threshold.

To formalize these concepts, denote $\theta_k(C)$ the optimal value of Model (2) which is the overall performance score for observation k evaluated based on criteria set C. The marginal impact of criterion $c \in C$, ρ_k^c, is measured as:

$$\rho_k^c = 1 - \frac{\theta_k(C\backslash\{c\})}{\theta_k(C)}, \tag{3}$$

where $\theta_k(C\backslash\{c\})$ is the score according to a *reduced* model, without criterion c. ρ_k^c is the percentage performance change due to the presence of c, and is referred to c's marginal impact on k. It is clear that $\theta_k(C\backslash\{c\}) \leq \theta_k(C)$. $\rho_k^c = 0$ $(\theta_k(C\backslash\{c\}) = \theta_k(C))$ indicates that criterion c's marginal impact on k is zero; larger ρ_k^c suggests that criterion c has a more significant marginal impact on k.

If c is not a relevant criterion, we should observe that the performance is not substantially affected by the presence of c in the model. Pastor et al. (2002) propose a statistical approach to formalize this concept as: assume ρ_k^c are observed values of the random sample Γ_k, $k \in K$ is drawn from a population (Γ, F), Γ being randomly distributed according to F, where F is a cumulative density function $[1, \infty)$. The marginal impact of significance greater than an

individual impact threshold $\bar{\rho}$ is an event with probability $p = P\{\Gamma > \bar{\rho}\}$. If the probability p is high, it is very likely that criterion c plays an important role in Model (2) given $\bar{\rho}$. In fact, the probability p can also be interpreted as the proportion of the underlying population being affected. Denote T_k^c an indicating random variable as follows:

$$T_k^c = \begin{cases} 1 & \text{if } \Gamma_k > \bar{\rho}, \\ 0 & \text{otherwise,} \end{cases} \quad k \in K, c \in C. \tag{4}$$

Then T_k^c is an indicator of whether c affects k, and follows a Bernoulli distribution with parameter p. Given T_k^c, $k \in K$ distributed Bernoulli(p), $T^c = \sum_{k \in K} T_k^c$ is the total number of observations affected by the presence of c, and follows binomial $(|K|, p)$. The following hypothesis tests the relevance of c:

$$H_0\colon p \leq p_0 \quad \text{against} \quad H_A\colon p > p_0. \tag{5}$$

Rejecting H_0 in (5) suggests criterion c is significant in the performance evaluation of more than $p_0 \times 100\%$ of observations because their overall performance scores are affected by more than $\bar{\rho} \times 100\%$ when c is not in the model. Therefore c should be considered as a relevant criterion in Model (2). Test (5) is a standard proportion hypothesis test, and a simple test based on binomial $(|K|, p)$ can be used once p's are obtained. We note that the correctness of the statistical approach is based on some important assumptions that can be validated by Monte Carlo simulations (Pastor et al. 2002).

Test (5) provides a statistical means to determine whether a particular criterion is relevant in the performance evaluation model. The thresholds on marginal impact $\bar{\rho}$ and probability of being affected p_0 must be specified prior to executing the procedure. We apply a backward elimination procedure to determine the proper criteria starting with a full set of all criteria candidates. In each iteration all criteria are tested for significance using (5), and the least significant one will be removed. Rejecting H_0 concludes that the criterion should remain. For the criteria for which the hypothesis cannot be rejected, we remove the one having the smallest number of observations that are affected beyond the marginal threshold since it is clearly the least significant (see Fig. 1 for the pseudocode). After a criterion is removed the next iteration begins. The procedure stops when the hypothesis test (5) is rejected for all criteria or when only one minimizing and one maximizing criteria remain.

Time windows, comparing several years of data in a single analysis, can be used to increase the robustness of the model specification approach. We note that the length of the time period is arbitrary; there are trade-offs when choosing time window length. Due to the data distribution, a shorter time window may cause more turbulence in the analytical results, particularly when using single-year analysis. A longer time window reduces the influence of data variation but may be less sensitive in identifying new performance criteria. In addition we observe that similar effects can be achieved by adjusting $\bar{\rho}$, individual impact threshold, or p_0, the population impact threshold.

The proposed method is iteratively applied to data in different periods to determine the proper criteria set for the corresponding periods. While the performance criteria are determined independently for each period, the data is generated by pitchers and managers that compete year after year. Thus the criteria selection results over time reveal the dynamics of the performance space and the changes in behavior of the pitchers and the managers. In this paper the changes are interpreted to reflect the change in strategy of the team as led by the manager.

```
C = I ∪ O :     I and O are full input and output set respectively.
K :             Set of observations

procedure removeC(C, K)
{
    for k ∈ K  {compute θ_k(C) by (2);}

    for c ∈ C  {
        T^c = 0
        for k ∈ K  {
            compute θ_k(C \ {c});
            if (θ_k(C \ {c}) and θ_k(C) are significantly different, i.e., ρ_k^c ≥ ρ̄ ) {
                T^c = T^c + 1
            }
        }
    }

    find argmin_c T^c
    if (T^c / |K| is statistically insignificant for p_0 ) {
        // i.e., conclude p is smaller than p_0
        C = C \ {c}
    } else terminate the loop
}
```

Fig. 1 Pseudocode for removing one criterion

4 Case study: MLB pitchers

MLB pitching performance records exist for the 135 years from 1871 to 2006.[1] Obviously, the number of teams and pitchers vary from year to year. Figures 2 and 3 illustrate the changes in total number of teams and pitchers used in our database; total pitchers range from 17 (1874) to 700 pitchers (2006). One reason for the increase in total pitchers is the growth in the number of teams; there are less than 10 teams on average in the first decade of the database and 30 teams after the sport's expansion in 1998 (Fig. 2). Figures 2 and 3 comparing teams and pitchers show 33 teams with only 236 pitchers (1884) and 30 teams with 700 pitchers (2006).

Baseball's expansion is not the only reason for using more pitchers. Figure 4 presents the changes in the number of games played each year for each team (seasons prior to 1884 have less than 100 games per season; 1981, 1994 and 1995 are short seasons due to labor strikes). Clearly, the total number of pitchers used is also related to roster size, team strategy, quality of opposing pitchers, and availability of talent.

[1] http://www.baseball-databank.org/

Fig. 2 Number of teams

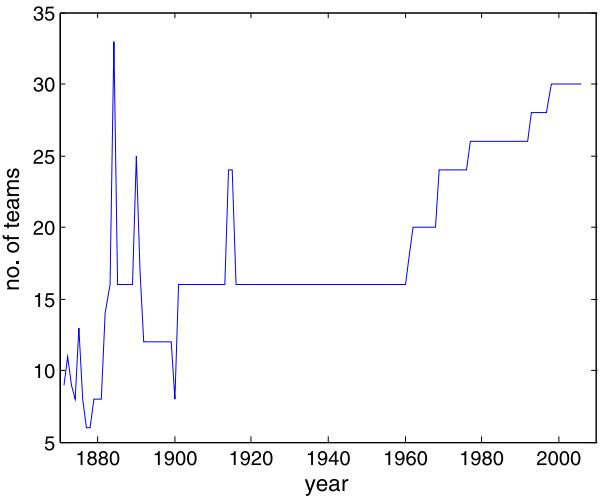
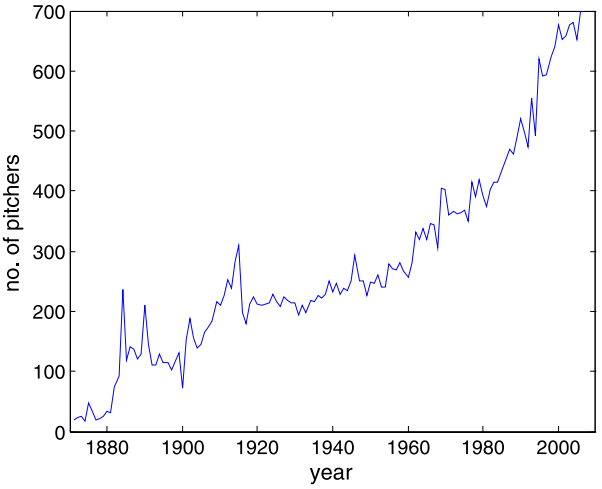


Fig. 3 Number of pitchers



The parameters of our approach are specified on a percentage basis in order to maintain robust results over the fluctuations in the number of pitchers. To analyze pitcher performance (using the method described in Sect. 3), we select nine major performance criteria representing different aspects of an individual pitcher's performance as the full set of criteria C. They are: games (G); earned runs (ER); outs pitched—innings pitched $\times 3$—(IPOuts); hits allowed (H); wins (W); shutouts (SHO); strikeouts (SO); saves (SV); and complete games (CG). G, ER, IPOuts and H are criteria to be minimized, $I = \{\text{G, ER, IPOuts, H}\}$. W, SHO, SO, SV and CG are to be maximized, $O = \{\text{W, SHO, SO, SV, CG}\}$.

5 Analysis, results and interpretations

The following discussions address the implementation of our approach for the MLB data set and demonstrate how our proposed approach reveals the dynamics of the performance

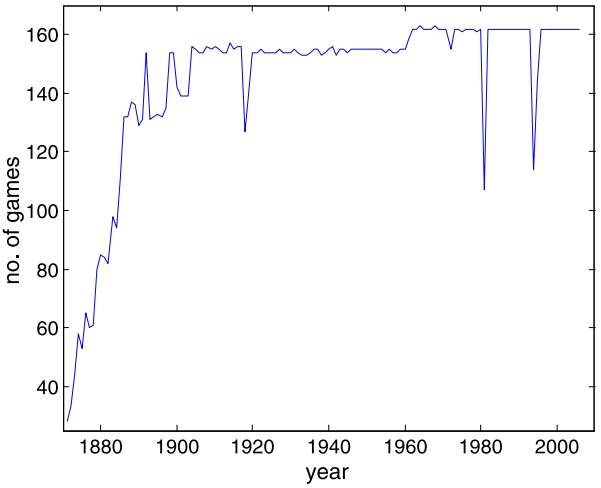


Fig. 4 Games played each year for each team

Table 1 Criteria inclusion frequency (individual impact threshold $\bar{\rho} = 5\%$, population impact threshold $p_0 = 7.5\%$)

Criterion	ER	G	H	IPouts	CG	SHO	SO	SV	W
# being selected*	1	132	0	130	100	2	109	77	109

*Out of 132 periods

space for evaluating the baseball pitchers. The findings provide several insights about the identification of strategy changes.

We select a five-year overlapped time window comprising records that are pooled in clusters of five consecutive years for a total of 132 five-year periods. The time stamp for each record is the last year of the five consecutive years (e.g., year 2006 represents data from 2002 to 2006). We report results related to the DEA specification in Banker et al. (1984). The benefit of using this model is to compare performance among pitcher with similar number of appearances. In the initial analysis we use $\bar{\rho} = 0.05$ and $p_0 = 0.075$ and statistical significance level $\alpha = 0.05$ for test (5). As with any statistical test, the thresholds selected are arbitrary and the trade-offs exist. However, typically relatively low values are used so that the development of new strategies will be detected. We also consider several different values for the window size, $\bar{\rho}$, and p_0, and additional results are available upon request.

Table 1 presents the number of periods (out of 132) in which each of the nine criteria is selected for use in the performance evaluation Model (2). Table 1 shows that H (hits) is never selected while ER (earned runs) is only selected once[2] and SHO (shutouts) is only selected twice[3] in the performance evaluation model. These results indicate that other measures of performance account for pitchers' performance, thus hits, earned runs and shutouts are not necessary in the model. In contrast G (games) is selected for all 132 periods, and IPouts

[2] 1992.

[3] 1925 and 1926.

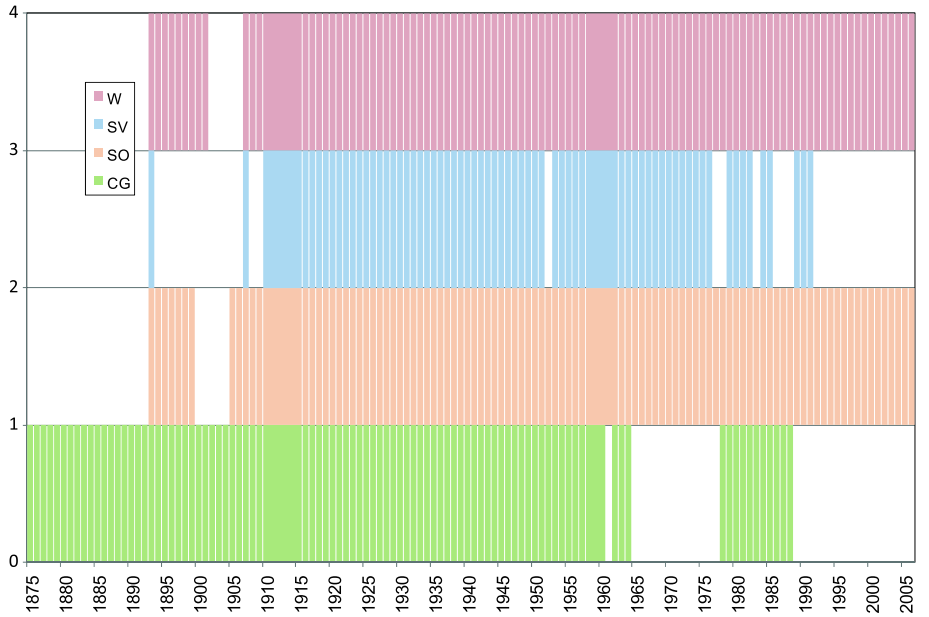


Fig. 5 Inclusion of criteria (individual impact threshold $\bar{\rho} = 5\%$, population impact threshold $p_0 = 7.5\%$)

(innings pitched) is excluded twice.[4] These results indicate that games and innings pitched are more important to quantify the pitcher's opportunities and thus important to include in the performance measure. From this we may conclude that it is of little concern if the pitcher gives up a few runs or allows a few hits since the manager's (principal) goal (winning) can still be met. We observe that shutouts are a weak differentiator of pitchers' performance because it is not necessary to hold the other team scoreless to achieve the goal of winning. Although it may be clear that these variables are not necessary to quantify performance, this analysis is a statistical validation that the variability in the criteria causes their relationship with the pitchers' overall performance (as defined by the criteria that remain in the model) to be unimportant.

Complete games, strikeouts, saves and wins that appear in the model as criteria periodically over the analyzed time horizon (Table 1) reveal the dynamics of the performance space. A detailed analysis of the four provides further insight into information regarding the principal's underlying strategy. Figure 5 shows the periods between 1875 and 2006 in which W, SV, SO and CG (from the top to the bottom) are included in the performance evaluation model.

W (wins) is perhaps the most interesting measure of pitchers' performance. If winning is the goal for both the team and the manager, it is also the objective of any strategy the manager implements. However, W[5] is a weak differentiator of a pitcher's performance in 23

[4] 1887 and 1888.

[5] A pitcher receives credit for a win when he/she is the team's pitcher at the time that the team takes a lead and keeps it for the remainder of the game. A pitcher who starts a game but does not pitch at least five full innings will not be credited with a win. Instead, the win is credited to the first relief pitcher entering the game if the team it is able to maintain the lead. The winning pitcher cannot be credited with a save in the same game.

years due to the coaching strategy and batting performance of the pitcher's team. A pitcher who pitches well by all accounts, but does not receive run support from teammates will win few games. Further, if the team has talented pitchers in its bullpen the manager may be willing to use one as a relief pitcher when the game is tied and a few batters reach base. This scenario contrasts with the more popular strategy historically of allowing the starting pitcher to "work through a jam".

Regarding W, another historical strategy of significance in the late 1800's and early 1900's is to keep a pitcher in the game when there is no cause for removal. Between 1875 and 1905 a high percentage of games started were completed and only when a pitcher was performing very poorly were they removed; thus the complete game criterion was more informative during this period. And much of the performance that could be capture by W was already being explained by complete games. This reduces the significance of the W measure to the point where it is excluded from the performance model in many years between 1871 and 1905.

SO (strikeouts) tell a similar story. SO are included as an important performance measure in all but 23 periods (most of which occurred before 1905 (Fig. 5)). The strategy of the power pitcher who does not allow opposing batters to hit the ball was not commonly used before 1900. Thus strikeouts were largely a function of innings pitched and did not give a good indication of the quality of the pitchers' performance.

Figure 5 also shows that CG (complete games) is a key criterion in every period prior to the mid-1960's (later it makes a brief appearance in 1980's). In the early 1960's only 25% of the games started are completed and nearly two-thirds of starting pitchers complete at least one of their starts. The average pitcher who completes at least one game completes six games. However, there are only six pitchers above the average and 181 below the average (in 1963, for example), indicating that only a select few pitchers with good performance will be identified by this criterion. Each pitcher with more than six complete games also has more than ten wins, indicating that W may capture the indications of good performance available in complete games. Thus, even though CG is still an important criterion in the early 1960's, there are already signs that W may capture similar information along with providing information on the population of pitchers that do not complete any games.

SV (saves) are not observed as an important criterion before the 1910's. Historically, bullpens were not very deep and most managers' strategies did not include the employment/use of specialty relief pitchers (given the number of teams and games played in modern times, managers prefer to "save" pitchers' arms; it is quite common to see a star pitcher pulled from a game after five innings). In the early 1900's one manager introduces the use of a pitcher solely in save situations (Carminati 2004), but this new strategy is not clearly identified and adopted until the late 1950's. Thus, the use of relief pitchers earns recognition for a significant set of relief pitchers by crediting them with a save in retrospect; however, managers are generally unaware of the strategy.

SV are not adopted as an official MLB statistic until 1969, and are subsequently calculated for historical data. The strategy of an ace closer, now in common use, has grown in popularity since the early 1970's. An ace closer is a relief pitcher who is only called upon to pitch the ninth inning in a save situation (James 2003). As the number of pitchers has grown with the development of the set-up man strategy and other specialized relief pitchers and the ace closer strategy limiting the number of players earning SV has gained acceptance, the proportion of pitchers earning saves has fallen. Thus, for a fixed impact level (7.5%) and a growing population of pitchers, if every team employs the ace closer strategy, then SV will become insignificant because only one pitcher per team is earning saves and this may not be enough to reach the threshold. This of course is an approximation, because some teams

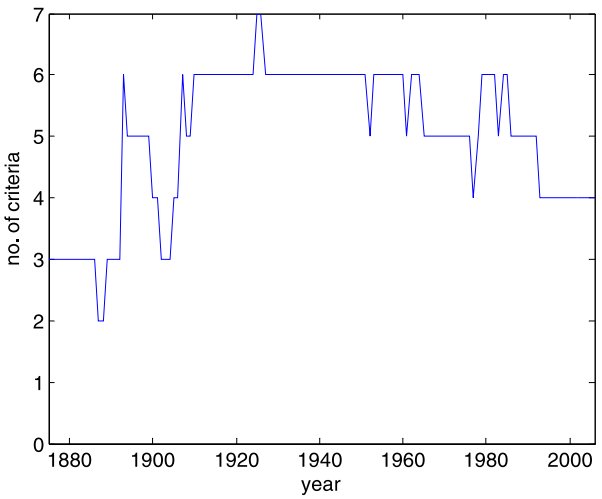


Fig. 6 Number of criteria included for each period

do not use the ace closer strategy, and injuries or rest schedules provide other pitchers with the opportunity to earn saves. However, this observation reveals that as the proportion of pitchers used in executing a particular strategy varies depending on strategy. Thus when a strategy such as an ace closer does not require many pitchers to implement the performance measure associated with the strategy may no longer be identified in the model even though the strategy is still used.

Figure 6 summarizes the change in the number of performance dimensions over the time period. The initial set of criteria use was developed in the early 1900's. Therefore, it is natural that the largest number of criteria is included between 1910 and 1990 in the model. As mentioned above since strategies used by principals evolve over time new performance measures are needed. For example, the use of an eighth and perhaps a seventh inning setup pitcher is becoming popular. However, the MLB has not yet officially adopted the statistic of "holds"[6] to quantify the performance of a pitcher in this role. The need for new performance measures is further indicated by the failure to include the current measures in models characterizing good performance (since 1993 only 4 of the 9 variables are important as shown in Fig. 6).

6 Parameter sensitivity analysis

As mentioned, thresholds are selected arbitrarily and the trade-offs exist. We present the sensitivity analysis of the population impact threshold in this section.

The analysis in Sect. 5 is based on population impact threshold $p_0 = 0.075$, which means selected criteria affect more than 7.5% of the population statistically. On the other hand, a criterion is unimportant statistically if less than 7.5% of the population is affected, (i.e. more than 92.5% of the population do not consider it to be important). Similarly, a criterion is unimportant when setting $p_0 = 0.05$ means that more than 95% of the population considers

[6]A relief pitcher coming into a game to protect a lead who gets at least 1 out and leaves without giving up that lead will be credited with a hold. But a reliever cannot be given a save and a hold simultaneously (Major League Baseball 2008).

Table 2 Criteria inclusion frequency (individual impact threshold $\bar{\rho} = 5\%$, population impact threshold $p_0 = 0.05$)

Criterion	ER	G	H	IPouts	CG	SHO	SO	SV	W
# being selected*	7	132	0	132	106	7	116	102	111

*Out of 132 periods

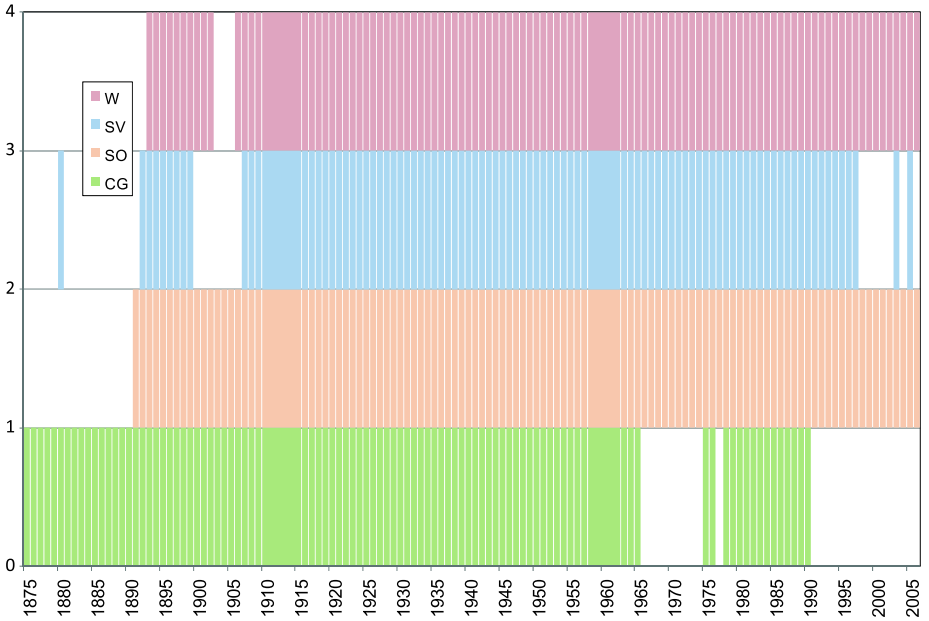


Fig. 7 Inclusion of criteria (individual impact threshold $\bar{\rho} = 5\%$, population impact threshold $p_0 = 5\%$)

this particular criterion irrelevant. Clearly $p_0 = 0.05$ leads to the inclusion of more criteria in Model (2). A criterion c selected to be in Model (2) at a particular time period with $p_0 = 0.05$, but not included in the model when $p_0 = 0.075$ may reveal the introduction of a new strategy that is not yet widely used. When criterion c is included in the analysis with the higher population impact threshold, it indicates an increase in the popularity of a successful strategy.

The sensitivity analysis on population threshold is applied for the MLB case. Table 2 reports results that are similar to Table 1; however the population threshold is $p_0 = 0.05$. Table 2 shows that more criteria are selected in each time period. Figure 7 is similar to Fig. 5 for the case when $p_0 = 0.05$. Figure 7 shows that SV is a relevant criterion consistently for 5% of the pitchers from 1907 to 1997. Similarly CG is a relevant criterion to evaluate pitchers' performance until 1990 (with a brief exception in 1977 and 1966–1974). The difference between Figs. 5 and 7 reveals that CG is a key criterion until 1965 when the popularity dropped off quickly. A fewer managers considered the strategy and it built in popularity during the late 1970's, but then again dropped off quickly in 1990 and has not been popular since.

Here a sensitivity analysis has been presented. Varying the p_0 parameter, the population impact threshold, gives an indication of the percentage of the population that has adopted a

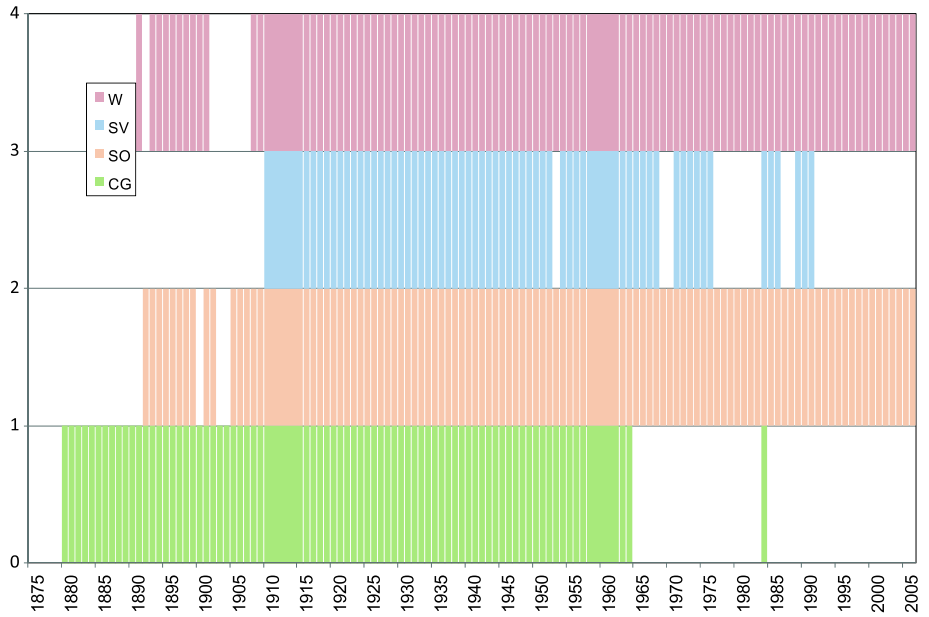


Fig. 8 Inclusion of criteria for NL pitchers (individual impact threshold $\bar{\rho} = 5\%$, population impact threshold $p_0 = 7.5\%$)

particular strategy. While the focus of this paper has been on using historical data to identify strategic changes, it is clear these tools have implications for the literature on innovation diffusion, see for example, Rogers (1962) or Alexander and Nelson (1973). Attempts to integrate these two literatures have been documented in work such as Kumar and Russell (2002) and Timmer and Los (2005); however, the further development of this integration is an area that has significant potential.

7 Modeling remarks

Relative performance evaluation depends on the selection of both criteria and the peer group. This work focuses on criteria selection given a defined peer group. Investigations of different peer groups can provide insight into the strategic differences of the groups, and may yield interesting directions for the future research.

We defined and studied the performance space for MLB pitchers, including both American and National leagues (AL and NL); the results represent the strategic dynamics of MLB. However we note that there are reasons to believe the two leagues may differ significantly from a pitching perspective, such as rules (e.g., designated hitter), umpire crews and ball parks, etc. Detailed investigations on NL and AL separately may reveal whether the strategy differences exist in two leagues.

We further analyze NL and AL separately; Figs. 8 and 9 reveal the dynamics for NL and AL,[7] respectively. Both figures show similar patterns to Fig. 5, and imply that the major

[7]NL data started in 1876 and the first analysis record is 1880. AL data started in 1901 and the first record is 1905.

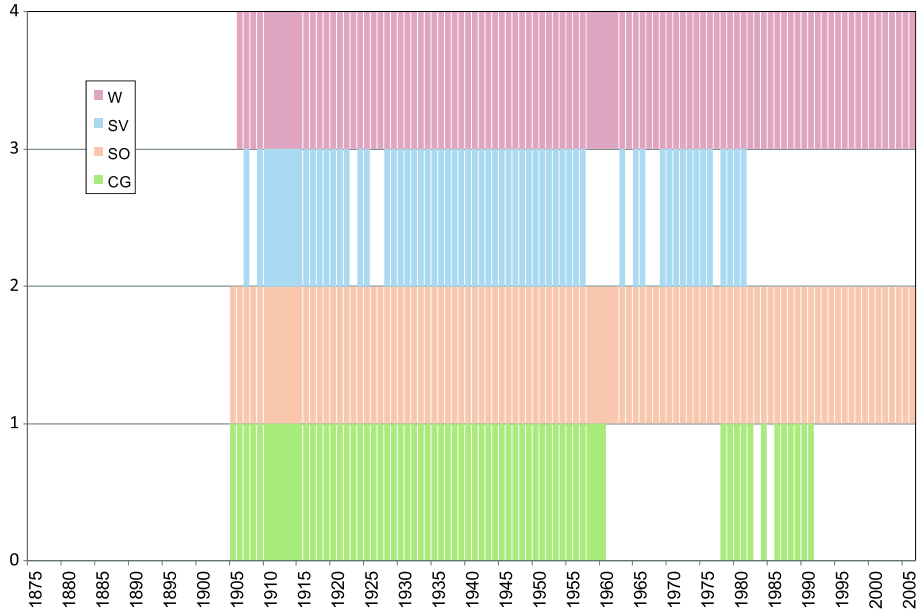


Fig. 9 Inclusion of criteria for AL pitchers (individual impact threshold $\bar{\rho} = 5\%$, population impact threshold $p_0 = 7.5\%$)

trends in strategy are similar. However, there are some differences worth comment. Saves is first included in the criterion for NL pitchers from 1910 and continued to be an important criteria through 1991. While, the use of relief pitchers was also adopted in the AL the importance and wide spread implementation was not as evident as in the NL. The ration may be related to the differences in rules between the two leagues. The American league has use of the DH, thus it is not necessary to remove the pitcher to improve offensive performance. In the NL the pitcher bats and is typically one of the weakest batters, thus it is often desirable to remove a pitcher regardless of his pitching performance in order to insert a stronger batter in the pitcher's batting position. These offensive types of substitutions impact the NL by reducing the likelihood of complete games and increasing the opportunities for saves. Both of these effects can be seen in Figs. 8 and 9. Note the instability of the evaluation criteria is due to threshold selection, p_0, relative to size of the population of pitchers used to adopt a particular strategy. As discussed in Sect. 6, by selection a lower threshold criteria a more stable criteria set could be achieved.

8 Conclusion

A central decision-maker (manager) uses performance evaluation to measure an agent's (pitcher's) contribution towards the ultimate goal(s) of a firm (team). The agents implement the strategies of the decision-maker and the results of the executions are observed and collected as the performance data. Because the strategies of the decision-maker change over time, the set of performance criteria spanning the performance space will also change. Using a DEA-based approach developed in Pastor et al. (2002), this paper extends the use of that model to consider the dynamics of the performance space in a panel data setting with

minimum subjective judgments and assumptions imposed on the data. The findings point to new directions for knowledge discovery, particularly in identifying revealed preferences or strategies based upon historical and modern observations. The proposed method was applied to MLB pitchers' performance evaluations from 1871 to 2006. The findings showed that the relevant criteria for evaluating pitchers' performances have changed significantly over time. The trends identified in the criteria are consistent with experts' opinions although they were objectively drawn from data. We observe that the proposed method does not intend to replace the experience and knowledge of experts. We view it as a complementary tool and a statistical validation depending on the availability of a large dataset. The proposed method is not limited to baseball data or time series analyses, but can be used to identify strategic differences due to factors, e.g., geographical locations or leagues as discussed in Sect. 7; however, this is a topic for future research.

Acknowledgement This research was supported in part by the National Science Council, Taiwan (NSC 96-2221-E-009-033).

References

Alexander, A. J., & Nelson, J. R. (1973). Measuring technological change: Aircraft turbine engines. *Technology Forecasting and Social Change*, *5*, 189–203.

Anderson, R. T., & Sharp, G. P. (1997). A new measure of baseball batters using DEA. *Annals of Operations Research*, *73*, 141–155.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiency in data envelopment analysis. *Management Science*, *30*(9), 1078–1092.

Carminati, M. (2004). A graphical history of relief pitching. http://www.baseballgraphs.com/main/index.php/site/article/a_graphical_history_of_relief_pitching/. Accessed August 5, 2008.

Casu, B., Shaw, D., & Thanassoulis, E. (2005). Using a group support system to aid input-output identification in DEA. *Journal of the Operational Research Society*, *56*(12), 1363–1372.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429–444.

Cinca, C. S., & Molinero, C. M. (2004). Selecting DEA specifications and ranking units via PCA. *Journal of the Operational Research Society*, *55*, 521–528.

Cook, W. D., & Zhu, J. (2007). Classifying inputs and outputs in data envelopment analysis. *European Journal of Operational Research*, *180*(2), 692–699.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, *132*, 245–259.

Golany, B., & Roll, Y. (1989). An application procedure for DEA. *Omega*, *17*(3), 237–415.

Hadley, L., & Ruggiero, J. (2006). Final-offer arbitration in major league baseball: A nonparametric analysis. *Annals of Operations Research*, *145*, 201–209.

James, B. (2003). *The new Bill James historical baseball abstract*. Glencoe: Free Press.

Kumar, S., & Russell, R. R. (2002). Technological change, technological catch-up, and capital deepening: Relative contributions to growth and convergence. *American Economic Review*, *92*(3), 527–548.

Lewis, H. F., & Sexton, T. R. (2004). Data envelopment analysis with reverse inputs and outputs. *Journal of Productivity Analysis*, *21*(2), 113–132.

Lovell, C. A. K., & Pastor, J. T. (1997). Target setting: An application to the branch network of a bank. *European Journal of Operational Research*, *98*, 290–299.

Major League Baseball (2008). MLB Miscellany: Rules, regulations and statistics. http://mlb.mlb.com/mlb/official_info/about_mlb/rules_regulations.jsp. Accessed August 11, 2008.

Pastor, J. T., Ruiz, J. L., & Sirvent, I. (2002). A statistical test for nested radial DEA models. *Operations Research*, *50*(4), 728–735.

Pastor, J. T., Lovell, C. A. K., & Tulkens, H. (2006). Evaluating the financial performance of bank branches. *Annals of Operations Research*, *145*, 321–337.

Rogers, E. M. (1962). *Diffusion of innovations*. Glencoe: Free Press.

Simar, L., & Wilson, P. W. (2008). Statistical inference in nonparametric frontier models: Recent developments and perspectives. In H. O. Fried, C. A. K. Lovell, & S. S. Schmidt (Eds.), *The measurement of productive efficiency and productivity growth*. London: Oxford University Press.

Timmer, M. P., & Los, B. (2005). Localized innovation and productivity growth in Asia: An intertemporal DEA approach. *Journal of Productivity Analysis*, *23*(1), 47–64.

Wanger, J. M., & Shimshak, D. G. (2007). Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives. *European Journal of Operational Research*, *180*, 57–67.