Economic prediction of sport performances from the Beijing Olympics to the 2010 FIFA World Cup in South Africa: the notion of surprising sporting outcomes

Wladimir Andreff, Madeleine Andreff

In: Placido Rodriguez, Stephan Késenne & Ruud Koning, eds., *The Economics of Competitive Sport*, Edward Elgar, Cheltenham 2015 (forthcoming):

CHAPTER 11

ECONOMIC PREDICTION OF SPORT PERFORMANCES FROM BEIJING OLYMPICS TO 2010 FIFA WORLD CUP IN SOUTH AFRICA: THE NOTION OF SURPRISING SPORTING OUTCOME

Wladimir Andreff * **& Madeleine Andreff** **

* Professor Emeritus at the University of Paris 1 Panthéon Sorbonne, Honorary President of the International Association of Sport Economists, andreff@club-internet.fr

** Former Senior Lecturer in Statistics and Econometrics at the University of Marne-la-Vallée, madandreff@gmail.com

Introduction

The distribution of medal wins across nations at Summer Olympics is extremely uneven between developed and developing countries: the former – about 40 nations – concentrate from two-thirds to three-quarters of medal wins while the latter - about 160 nations – obtain from one-quarter to one-third of medals total. This observation suggests a likely relationship between a nation's Olympic sport performance and its level of economic development. Indeed, it has been empirically verified that the number of medals a country wins at Summer Olympics significantly depends on its population and GDP per inhabitant (Andreff, 2001). Thus, in a sense, the number of medal wins at the Olympics can be regarded as an additional index of economic development, just like the literacy rate, the percentage of children attending primary school, health expenditures per inhabitant, mortality or morbidity rates. On the other hand, the level of economic development and population could be used as realistic predictors of Olympic performances.

The only sport mega-event which can compare to Summer Olympics in terms of fan attendance, TV viewing and economic impact is FIFA soccer World Cup. Nobody knows whether a nation's level of economic development may impact on its sport performance at soccer World Cup since such an issue remains unheeded in the literature so far, whatever one refers to sports economics or development economics. One motivation of this article is to provide a first insight into this issue and, the other way round, check whether a nation's performance at soccer World Cup may have any sense in reflecting its economic development. Econometric estimation of how

much significant are the economic determinants of medal wins by each participating nation is now quite usual. Our core research question is: would a model based on population and GDP per capita as determinants perform as well in explaining soccer World Cup outcomes as it is used to perform with Olympic medal wins? Since the estimated model has provided a good prediction of medal wins at the next Olympics, would it be able to predict FIFA World Cup outcomes with a similar success?

After a brief look at modelling and predicting Summer Olympics medal distribution (**1**), a slightly improved model is estimated and then implemented to predict how many medals each nation would have obtained at the 2008 Olympics (**2**). The prediction is compared to actual outcomes observed in Beijing (**3**). A next step is to understand why a similar prediction model has not yet emerged with regards to FIFA World Cup: a major reason is that the soccer World Cup outcome is rather unpredictable due to a number of "surprises" – surprising outcomes – during its final tournament (**4**). Thus, adapting the model of Olympic medals prediction to FIFA World Cup requires that some football-specific or "footballistic" variables be introduced alongside with economic variables (**5**). Such emended model is estimated on the basis of past FIFA World Cup results (**6**), and then used to predict the semi-finalists at the 2010 World Cup in South Africa (**7**). Since the predictions regarding the last soccer World Cup do not exhibit good results, performance in the latter is meaningless as an index of economic development. Moreover, this opens an avenue for further research about the notion of surprising sporting outcome and its metrics (**8**). The conclusion emphasizes that economic prediction of sport performances is to be taken with a pinch of salt.

1. Economic determinants of Olympic medals

More than thirty studies have looked for the determinants of Olympic performances since 1956 combining socio-economic variables with weather, nutrition, mortality in the athlete's home nation, protein consumption, religion, colonial past, newspapers supply, urban population, life expectancy, geographical surface area, military expenditures, judicial system and those sport disciplines taught at school. A widespread assumption across sports economists who have participated to these studies is that a nation's Olympic performance must be determined by its endowment in – and the level of development of - economic and human resources captured through GDP per capita and population. Notice that an increase in the number of medal wins by one country logically is an equivalent decrease in medals won by all other participating nations. Therefore, if one wants to explain the Olympic performance of one specific nation, one has to take into account all other participating nations within the overall constraint of the distributed medals total.

During the cold war period, another significant variable emerged: a nation's political regime. The first Western work attempting to explain medal wins by nations' political regime (Ball, 1972) triggered a Soviet rejoinder (Novikov and Maximenko, 1972), both differentiating capitalist from communist regimes. The first two econometric analyses of Olympic Games (Grimes *et al.*, 1974; Levine, 1974) exhibited that, when regressing medal wins on GDP per capita and population, communist countries were outliers: they were winning more medals than their level of economic development and population were likely to predict. A last variable has been introduced, namely since Clarke (2000), which is the influence on medal wins of being the Olympics hosting country, *i.e.* a sort of home advantage. The host gains more medals than otherwise due to big crowds of national fans, a stronger national athletes' motivation

when competing on their home ground and being adapted to local weather, and not tired by a long pre-Games travel.

Econometric methodology has developed in more recent studies such as an ordered Logit model (Andreff, 2001), a Probit model (Nevill *et al.*, 2002) or an ordered Probit model (Johnson and Ali, 2004) in which a quadratic specification in GDP per capita is employed to capture a postulated inverted U-shaped relationship meaning that higher levels of GDP per capita have a positive impact on medal wins though decreasing after some threshold. The most quoted reference is Bernard and Busse (2004) whose model has been widely used in further studies. In this Tobit model implemented for estimating and predicting Olympic performances, the two major independent variables - GDP per capita and population – are taken on board with three dummies that capture a host country effect, the influence of belonging to Soviet-type and other communist (and post-Soviet and post-communist after 1990) countries as against being a capitalist market economy.

2. Predicting Olympic medals distribution in Beijing 2008

Starting from Bernard and Busse, after a few emendations, a more specified model has been elaborated on (Andreff *et al.*, 2008). The dependent variable is each nation's number of medal wins[1]: $M_{i,t}$. The first two independent variables are GDP per capita and population. Contrary to Bernard and Busse, it is not assumed here that preparing an Olympic team is timeless and, then, independent variables are four years lagged: GDP per inhabitant $(Y/N)_{i,t-4}$ in 1995 purchasing power parity dollars and population $N_{i,t-4}$ (World Bank data). The assumption is that four years are

required to build up, train, prepare and make an Olympic team the most competitive in due time. A *Host* dummy is used to capture a home advantage.

Bernard and Busse divide the world into communist regimes and capitalist market economies which obviously fits with the cold war period. Since then, this is too crude with regards to post-communist transition economies: the sports economy sector has differentiated a lot across former socialist countries during their institutional transformation process (Poupaux and Andreff, 2007). Such differentiation has translated into a scattered efficiency in winning Olympic medals after 1991 (Rathke and Woitek, 2008). A first emendation to Bernard and Busse's model is introduced here with a classification that distinguishes Central Eastern European countries (*CEEC*) which have left a Soviet-type planned economy in 1989 or 1990, and transformed into a democratic political regime running a market economy: Bulgaria, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia (and Czechoslovakia until the 1993 split), Slovenia, and the GDR (until German reunification in 1990). Another commonality to this group is that these countries have joined the European Union.

A second country group (*TRANS*) gathers new independent states (former Soviet republics) and some former CMEA member states which have started up a similar process of transition but are lagging behind the CEECs in terms of transformation into a democratic regime and are stalling on the path toward a market economy: Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Moldova, Mongolia, Russia, Tajikistan, Turkmenistan, Ukraine, Uzbekistan and Vietnam. None of them has joined the EU so far or has really an option to do so. The next two groups have not been Soviet regimes properly speaking in the past, although they have been both communist regimes and planned economies. In the first one (*NSCOM*), we sample

those countries which have started up a transition process in the 1990s: Albania, Bosnia-Herzegovina, China, Croatia, Laos, Macedonia, Montenegro, and Serbia (and the former FSR Yugoslavia before the 1991 break up). Two countries have not yet engaged into a democratic transformation and a market economy: Cuba and North Korea. They must be considered as still communist regimes (*COM*). All other countries are regarded as capitalist market economies (*CAPME*), the reference group in our estimations. Table 1 exhibits uneven medal distribution by political regime.

Insert Table 1 about here

Beyond Bernard and Busse, a variable supposed to capture the influence on Olympic performance of a specific sporting culture in a region is introduced. For example, Afghan (and other Middle East) ladies are not used to have much sport participation or attend sport shows, even less to be enrolled in an Olympic team. Resulting from these cultural disparities, some nations are more specialised in one specific sport discipline such as weight-lifting in Bulgaria, Turkey, Armenia, and the Balkans, marathon and long distance run in Ethiopia and Kenya, cycling in Belgium and the Netherlands, table tennis, judo and martial arts in various Asian countries, sprint in Caribbean islands and the U.S., and so on. It is not easy to design a variable that would exactly capture such regional sporting culture differences[2], but it is assumed that regional dummies may reflect them. The world is divided into nine sporting culture regions: *AFS*, sub-Sahara African countries; *AFN*: North African countries; *NAM*, North America; *LSA*, Latin and South America; *EAST*, Eastern Europe; *WEU*, Western Europe (taken as the reference region in our estimation); *OCE*, Oceania; *MNE*, Middle East; and *ASI*, (other) Asian countries.

Insert Table 2 about here

A first specification is simply *à la* Bernard and Busse, but with a differently defined political regime variable, with a censored Tobit model since a non negligible number of countries that participate to the Olympics do not win any medal. Therefore, a zero value of the $M_{i,t}$ dependent variable does not mean that a country has not participated and we work out a simple Tobit, not a Tobit 2 (with a two stage Heckman procedure)[3]. Dummies test whether the Olympic year is significant, taking 2004 as the reference. These dummies come out to be non significant. In a second specification, a data panel Tobit is adopted, in order to take into account unobserved heterogeneity, whose test is significant[4], and thus estimation with random effects is opted for. Data[5] encompass all Summer Olympics from 1976 to 2004, except 1980 and 1984 which are skipped out due to boycotts which have distorted the medal distribution per country. Therefore a first specification (**1**) is:

$$M^{*}_{i,t} = c + \alpha \ln N_{i,t-4} + \beta \ln \left(\frac{Y}{N} \right)_{i,t-4} + \gamma \, Host_{i,t} + \sum_{p} \delta_{p} \, Political \quad Re\,gime_{p,i} + \sum_{q} \kappa_{q} \, Year_{q,i} + \varepsilon_{i,t}$$

where $\varepsilon_{i,t} \sim N(0,\sigma^2)$

$M_{i,t}$ observation is defined by $M_{i,t} = \begin{cases} M^{*}_{i,t} \; if \quad M^{*}_{i,t} > 0 \\ 0 \quad if \quad M^{*}_{i,t} \le 0 \end{cases}$

A second specification (**2**) adds the above described dummy standing for sporting culture regions ($Region_{r,i}$):

$$M^{*}_{i,t} = c + \alpha \ln N_{i,t-4} + \beta \ln \left(\frac{Y}{N} \right)_{i,t-4} + \gamma \, Host_{i,t} + \sum_{p} \delta_{p} \, Political \quad Re\,gime_{p,i} + \sum_{r} \rho_{r} \, Re\,gions_{r,i} + u_{i} + \varepsilon_{i,t}$$

where $\varepsilon_{i,t} \sim N(0,\sigma^2_{\varepsilon})$ and $u_i \sim N(0,\sigma^2_{u})$

$M_{i,t}$ observation is defined by $M_{i,t} = \begin{cases} M^{*}_{i,t} \; if \quad M^{*}_{i,t} > 0 \\ 0 \quad if \quad M^{*}_{i,t} \le 0 \end{cases}$

A third specification (**3**) contains an additional variable $M_{i,t-4}$ on the right-hand side just like in Bernard and Busse who do not comment why they proceed in such a way. The interpretation here is that winning medals at previous Olympics matters for an Olympic national team which usually expects and attempts to achieve at least as well as four years ago. Such inertial effect is all the more relevant for those nations eager to win as many medals as possible, that is for most nations winning more than zero medals.

The third specification (**3**) is used to predict the medal distribution at Beijing Olympics:

$$M^{*}_{i,t} = c + \alpha \ln N_{i,t-4} + \beta \ln \left(\frac{Y}{N} \right)_{i,t-4} + \gamma \, Host_{i,t} + \sum_{p} \delta_{p} \, Political \quad Re\,gime_{p,i} + \sum_{r} \rho_{r} \, Regions_{r,i} + \theta \, M_{i,t-4} + \varepsilon_{i,t}$$

where $\varepsilon_{i,t} \sim N(0,\sigma^2)$

$M_{i,t}$ observation is defined by $M_{i,t} = \begin{cases} M^{*}_{i,t} \; if \quad M^{*}_{i,t} > 0 \\ 0 \quad if \quad M^{*}_{i,t} \leq 0 \end{cases}$

Insert Table 3 about here

All estimations deliver significant results (Table 3). In the first one, all coefficients are positive and significant at a 1% threshold, except for year dummies. It is confirmed once again that medal wins are determined by GDP per capita, population and a host country effect. Political regime is also an explanatory variable. The second estimation all in all exhibits the same results. The coefficients of regional sporting culture are significant except for Latin America, an area where the North American sporting culture may have permeated namely through Caribbean countries and Mexico (classified in *NAM*).

Since Western Europe is the reference a significant coefficient with a positive (negative) sign means that a region performs relatively better (worse) than Western Europe in terms of medal wins. Sub-Sahara Africa, North America and Oceania perform better. Though a little bit surprising for Sub-Sahara African countries since they are among the least developed in the world (except South Africa), this is due to a few countries which are extremely specialised in one sport discipline where they are capable of a non negligible number of medal wins, such as Ethiopia and Kenya in long distance runs. With negative coefficients, North Africa, Asia, Eastern Europe and Middle East perform worse. It is not surprising for North Africa and the Middle East due to some sport practice restrictions in the culture of various countries. In the case of Asia, only few countries are capable of a significant number of medal wins (China, both Koreas, Mongolia) given their GDP per capita. Since Eastern European countries are known as outliers - over-performers (given their GDP per capita and population) a negative coefficient results from the *Political Regime* already capturing their over-performance.

The pooling estimation[6] of Model 3 may suffer from endogeneity since the results may be biased by a correlation between the lagged endogenous variable and the error term. This issue is treated with a GMM dynamic panel (Arellano and Bond, 1991), a technique which provides estimated coefficients and predictions that are robust and close to those estimated with a Tobit model. Our predictions show up in Table 4 for a country sub-sample[7].

Insert Table 4 about here

The predicted first-rank winner is the United States, followed by Russia, and China which benefits from home advantage. Most developed market economies are

predicted to be among the major medal winners together with some post-communist transition countries.

3. Predictions and actual results: medal wins are rather predictable

Comparing predictions with the actual medal distribution that has come out from Beijing Olympics, the model performs well. It has provided good predictions: 70% of the observed results are encompassed in our predicted confidence interval (among those 189 countries for which data were available and computable). If prediction is assessed as acceptable when the error margin is not bigger than a two medal difference between prevision and actual outcome, then the model correctly predicts 88% of all Beijing results. The remaining unforeseen 12% account for surprises – unexpectedly surprising outcomes. The model correctly predicts the first ten medal winners, except Japan (instead of Ukraine), misses four out of the first twenty winners, though with a slightly different ranking. However, the most interesting is when model prediction is clearly wrong, that is basically for 23 nations, meaning that the five variables (plus the inertial variable) have not captured some core explanation of the Olympics outcome. Fortunately, economists are not capable to predict all Olympic results, otherwise why still convene the Games?

The major surprise in the actual outcomes, compared to model predictions, is the quite bigger than expected medal wins by the Chinese team – all published predictions have been wrong in this respect. The host country effect in China has been underestimated. Possibly, Chinese performance has also been boosted by

some undetected doping[8]. A second surprise is the underperformance of the Russian Olympic team, the worst since the cold war. Vladimir Putin convened the highest decision makers of Russian sport to command a new Olympic policy likely to avoid a repeated disaster at the 2012 London Olympics. In the same vein, some other transition countries, namely Romania, have won fewer medals than predicted in Beijing. The current state of restructuring the whole sports sector in these countries has not been sufficiently captured by our refined political regime variable.

The last three significant surprises are Great Britain, Jamaica and Kenya, the latter being the only two developing countries among the first twenty medal winners. Early preparation of a super-competitive team for the 2012 London Olympics may have been the cause for higher than predicted outcomes of the British team, as it is suggested by Maennig and Wellebrock (2008) who have introduced a "next Olympics host country" variable in their prediction. Great Britain's medals concentration in cycling (12 medals) may trace back again to undetected doping and/or deep specialisation of a nation in one sport discipline. The latter is the most likely explanation for Jamaican medals[9] concentrated in sprint and Kenyan medals in long distance runs. Though we have taken into account such specialisation through our lagged $M_{i,t-4}$ variable – Kenya had won 7 medals and Jamaica 5 in the same disciplines at Athens Olympics -, the inertia captured with this variable reveals to be insufficient.

4. Prediction of FIFA World Cup semi-finalists: why it is so hard?

The economics of FIFA World Cup outcome is much less developed than the economic approach to Olympic medal wins. There are two ways of explaining

international soccer successes in the literature. The most common method is to explain FIFA points and ranking (the FIFA/Coca Cola World Ranking for all national football teams) at one point in time (Hoffmann *et al.*, 2002b; Houston *et al.*, 2002; Macmillan and Smith, 2007; Leeds and Marikova Leeds, 2009; Yamamura, 2009). The second one consists in explaining a nation's success in FIFA World Cup over time. To the best of our knowledge, economic determinants of the soccer World Cup outcome have only been touched three times in the literature so far. From the three papers, it appears that surprising outcomes are the most common occurrence.

Torgler (2004) attempts explaining the determinants of the 2002 soccer World Cup outcome. The dependent variable is a dummy that measures whether a team wins a game or not in the World Cup final tournament. Explanatory variables are not economic. A variable captures the strength of a team through its FIFA ranking, and the positive influence on success of being the hosting team. A second set of variables is introduced regarding the performance of a team during a game: shots on goal, fouls, corner kicks, free kicks, off sides, cautions, expulsions, actual playing time (based on ball possession). The major result is that higher FIFA ranking leads to higher probability of winning a game: a one place improvement in world ranking increases a team's probability of winning by approximately 1%, but this result is not always significant. Higher number of shots on goal drives higher probability of winning; having a referee from the same region has a positive impact on the probability of winning a game, but this effect is not statistically significant[10].

A prediction model of FIFA World Cup outcome is due to Paul and Mitra (2008). It is not based on economic variables either. The authors remind that in the past four FIFA World Cup tournaments, 1994 to 2006, the top team in FIFA ranking never won, except Brazil in 1994. However, they test the relevance of the last FIFA ranking

published before the World Cup final tournament as a benchmark to evaluate teams' performance. In a Probit model, the dependent variable is a dummy that measures whether a team wins (1 = win, 0 otherwise) a game or not. The main explanatory variable is FIFA ranking with controlling for the number of goals scored by each team, and the number of yellow and red cards. A second OLS testing considers the scored goal difference as the dependent variable and FIFA rank difference is the main independent variable with controlling for goals scored, the number of yellow and red cards, the number of corner kicks, the number of fouls, the percentage of ball possession, and match attendance. Higher FIFA ranking is significantly associated with higher probability of winning a game. Higher-ranked teams score more goals. A more surprising result is that, though a higher number of yellow or red cards are less likely to win a game, in 2002 and 2006 teams with more yellow cards were more likely to win a game (and teams with more red cards in the 1998 Cup as well). Other surprises are that more corner kicks and more ball possession are associated with losing a game. Overall higher-ranked favourites have the winning trend in their favour, but there is a number of unexpected match outcomes. This is why it is so hard to estimate determinants and make predictions.

Monks and Husch (2009) test whether FIFA World Cup format may lead to a slightly rigged contest or, at least, whether it may favour certain teams, in particular the host country. In the tournament history, only seven teams have ever won the World Cup (Brazil 5 times, Italy 4, Germany 3, Argentina and Uruguay 2, England and France 1). Of the 18 tournaments held to date, the host has won six times. The authors test the impact of seeding, home continent and hosting on FIFA Cup outcome from 1982 to 2006. The dependent variable is a national team's World Cup final standing (from the winner down to the 32th among the qualified according to their performance during

the final tournament), and it is regressed on a team's FIFA rank before the World Cup, a dummy variable for being top seeded, a host country dummy, and a dummy variable if the World Cup is being played on a team's own continent. Ex ante rank is positive and significant in determining a team's final standing. Being top seeded results in an increase in final standing of approximately 5 places and the home continent advantage is approximately 2.8 places (but not significant). Both effects probably overlap with the host country variable (the host country is top seeded by definition) which provides 3 places better than the expected final standing, but the result is not significant. Rank, being the host country and playing on one's home continent[11] determine advancement in the tournament to either the quarterfinals or semi-finals.

5. Adapting the Olympics medal model to FIFA World Cup outcome

From the above-mentioned studies it is clear that explaining FIFA World Cup outcome is much harder than finding in socio-economic variables the determinants of Olympics medal wins, for different reasons. Soccer is a sport discipline which is more widespread in some countries (for instance Latin American countries) than others, whatever their level of economic development, the size of their population and their democratic or autocratic regime. Such specificity requires the introduction of some 'footballistic' variables in the estimation. To the contrary, the Olympics cover so many sport disciplines that overall economic development of a nation affects overall nation's sporting outcome, beyond disparities in performance across different sports – thus GDP and population are germane to stand for a significant share of the determinants. The number of surprising outcomes is much higher with the soccer

World Cup than with the Olympics also because in one case the surprises pertain to just one sport discipline whereas with the Olympics there are unexpected (surprising) medal wins in some sports that may, on average, compensate surprising medal "losses" in other sports for the Olympic teams from big (population) and rich (GDP per capita) nations.

Moreover, the two contests have different formats. In most Olympics disciplines[12], after a preliminary knock-out selection, eight athletes remain in contention for the finals and the first three best are rewarded with (gold, silver and bronze) medals during the finals. Thus it is not extremely tricky to build up an estimation of the determinants of medal wins - the first three ranked athletes (nations). It is more complex with FIFA World Cup final tournament since this contest combines a round robin first stage before the 1/8th finals and, then, a knock-out second stage from the 1/8th finals on. The uncertainty of outcome markedly increases from the first to the second stage (Monks and Husch, 2009) and, thus, the impact of economic variables might well dilute in the course of some knock-out games (thus the surprising outcome). This lays ground for the choice of dependent variable to have it as much comparable as possible with medal wins: it is chosen as the four nations making it for the semi-finals (*Semifin*) of a soccer World Cup final tournament. The determinants of being one of the four highest-ranked teams in the final tournament are looked for – and this facilitates using the same estimation model as the one explaining Olympics medal wins. The four highest-ranked are the winner, the finalist and two losing semi-finalists which play a ranking game the day before the final. Given the dependent variable (making for the semi-finals = 1; otherwise 0), a Probit model is estimated.

All national teams which have participated to the semi-finals are exhibited in Appendix 1 with their cumulative participation from the first 1930 World Cup up to

2006. Retaining the semi-finalists as the dependent variable also makes sense when referring to FIFA economic incentives. Given FIFA distribution rules, each team entering the World Cup final tournament earns a 3.79 million € bonus (in 2006). The next step – reaching the 1/8th finals – increases this amount by an extra 1.59 million €, followed by an additional 1.90 million € bonus when making it for the quarterfinals. Then for qualifying to the semi-finals, there is a huge jump of 6.33 million €, followed by only 630,000 € extra to make it for the finals and winning the finals adds another 1.27 million € (Coupé, 2007). In economic terms, it is rather significant to qualify for the semi-finals.

Independent variables are selected with a double purpose in mind: a/ comparing whether the same socio-economic variables play a role in determining FIFA World Cup outcome as with Summer Olympics medal wins; b/ finding a sample of socio-economic and footballistic variables that explain the soccer World Cup outcome in the long run, in order to come up with an ex post benchmark model that can be used further in ex ante predicting the semi-finalists of the 2010 World Cup. Due to data availability, the retained observation period runs from the 1962 soccer World Cup up to 2006, which includes 12 final tournaments. Data cover all national teams which have participated to soccer World Cup final tournaments since 1962 – that is 16 from 1962 to 1978, 24 teams from 1982 to 1994, and then 32 teams from 1998 on, *i.e.* 272 observations in an obviously unbalanced panel.

Population (*Pop*) and GDP per 1,000 inhabitants (*GDP/cap*) are the first two independent variables considered just like in the Olympics medal model (World Bank data). Squares are added for both variables (Pop^2 and GDP/cap^2), in tune with Houston *et al.* (2002) and Macmillan and Smith (2007), in order to control for possible decreasing returns of population and GDP per capita in terms of soccer World Cup

performance. The expectation is that population would have a positive effect on reaching the semi-finals while the specificity of soccer may lead to either significant or non significant effect of GDP per capita. These variables are introduced in the model with a two year time lag under a similar assumption as with the Olympics: the economic size and level of development of a nation two years ago is the context in which the preparation and training of a national soccer team starts up. In the two years after a FIFA World Cup, national teams are used to participate to a regional international contest such as UEFA Euro or the African Cup of Nations. Preparing the World Cup really starts up after the end of such contests (which means in *t-2*) when countries start playing the preliminary World Cup qualification stage at a regional level[13].

In previous studies, it has appeared that a nation's history in the football domain, such as World Cup appearances and the length of FIFA membership, matters when explaining its international soccer performance. Given our objective of explaining semi-finals participation, a specific semi-final history variable (*SFstory*) is derived from the data in Appendix 1. It is calculated by dividing all the figures in Appendix 1 by the number of FIFA World Cup final tournaments from 1930 up to the year appearing in a column of Appendix 1 (for instance, in the 2006 column, all figures are divided by 18, in 2002 by 17 and so on). This variable describes the uneven long-term capacity of a national team to make it for the semi-finals in a historical perspective and ranks nations according to this capacity. When one talks about "footballistic" nations or football-involved countries, Germany, Brazil, or Italy are often mentioned: indeed, they have been the most frequent semi-finalists at FIFA World Cups. As in previous studies, *FIFA rank* is tested as one explanatory variable, taking

FIFA ranking one month before the beginning of the final tournament, and a dummy (*Host*) for being the hosting country.

A regional variable (*Reg*) is different from the one used in the Olympics medal model. The latter's purpose was to capture a regional sport culture effect while in the case of FIFA World Cup it must measure the relative strength and density of elite football in six different geographical zones into which FIFA is divided, that is: AFC for Asia, CAF for Africa, CONMEBOL for South America, OFC for Oceania, UEFA for Europe, and CONCACAF for North, Central America, and the Caribbean. Seeding of the final tournament round robin stage varies across years but is based on teams' successes from each region in previous World Cups and organised in such a way as to assure that top-seeded teams will not have to play each other until the second phase (1/8th finals) of the final tournament (Monks and Husch, 2009).

A last assumption to be tested is whether a soccer-oriented nation, that is one which the number of players is relatively high compared to overall population, is successful in international soccer. The argument goes alongside with a pyramidal explanation of elite sport stating that the larger the mass of sport participants at the pyramid base, the better the elite top. Thus, most football-oriented nations should have highest probability to qualify for FIFA World Cup semi-finals. The number of (registered) soccer players (*Players*) divided by population can capture such possible effect.

Estimating the determinants of FIFA World Cup semi-finalists relies on a Probit model:

$$\Pr(Semifin^{*}_{i,t} = 1) = \Phi[a + b\,SFstory_{i,t-4} + c\,N_{i,t-2} + d\,N^{2}_{i,t-2} + e\left(\frac{Y}{N}\right)_{i,t-2} + f\left(\frac{Y}{N}\right)^{2}_{i,t-2} + g\,Host_{i,t} + h\,FIFArank_{i,t} + \sum_{r}\rho_{r}\,D_{r}Reg_{i} + k\,Players_{i,t}]$$

where Φ is the cumulative normal distribution.

The paucity of available data for *FIFArank* and *Players* has led to estimate three different specifications. FIFA ranking is only available since 1993, when FIFA started publishing it, whereas the number of registered soccer players in all national federations has been published only in 2000 and 2006 (FIFA Big Count, 2000 and 2006), which markedly reduces the size of the data sample. Thus in a first *M1* specification, these two variables are not taken on board. In a second *M2* specification, FIFA ranking is introduced but the sample is reduced to four World Cup final tournaments (1994 to 2006). Since FIFA ranking does not show up as statistically significant with *M2*, it is excluded in a third *M3* specification whereas the proportion of registered players in the population is taken on board, assuming that the data for 2000 is acceptable for estimating the 2002 FIFA World Cup outcome.

With a small and unbalanced panel, Probit estimation is used as a first step. Then to tackle the endogeneity of the semi-final history variable, a Probit model with an endogenous regressor and instrumental variables is resorted to. Valid instruments must be exogenous sources of variation in the semi-finalists, and it is difficult to think of candidate instruments relevant to explain international soccer performance (Macmillan and Smith, 2007). Thus, those exogenous variables of the best previous estimated model are retained as instruments.

6. Socio-economic and "footballistic" determinants of FIFA World Cup semi-finalists

Before estimating *M1*, a preliminary testing has shown that adding year dummies to *M1* comes out with none of these year dummies being significant. Therefore we do not proceed with panel data estimation.

Insert Table 5 about here

The estimation of *M1* shows that population and population squared is significant at a 1% threshold; the size of a nation matters with decreasing returns. Hosting the World Cup is also a significant determinant of making for the semi-finals. The host country has often muddled through the first round robin phase of the tournament to qualify for the semi-finals. The impact of belonging to each of the six regions on qualifying for the semi-finals is not significant for four regions out of six. Taking these four regions as the reference, Europe and South America show up as significant variables at a 1% threshold. Being a European or South American team significantly increases the probability of being semi-finalist. Most semi-finalists have been either European or South American teams so far. A last significant variable, though only at 10%, is the semi-final history variable. Having participated to past semi-finals has a positive effect on the probability of reaching this stage again. GDP per capita and its square are not significant. This makes a major difference between FIFA World Cup based on a single sport discipline and the multi-sport Olympics. The latter's outcome is determined by the level of economic development in participating countries whereas the former is not.

With *M2*, tested from 1994 to 2006, the introduction of FIFA ranking as an independent variable has a devastating effect. Most variables become non significant, namely population, population squared and hosting the World Cup. FIFA rank itself is not significant either. The problem with this variable is endogeneity since its calculation includes each team performance (namely qualifying for the semi-finals) in the past three World Cups[14] and thus FIFA ranking interferes with the semi-final history. The host country effect fades away from the determinants of qualifying for the semi-finals, against the frequent host nation expectation that its team has a home

advantage to qualify. Overall, *M2* is the most difficult specification to interpret even though it maintains the European and South American regions as significant determinants of making for the semi-finals. The semi-final history remains significant at 10% and prevails over FIFA ranking as the relevant footballistic variable.

The number of soccer players per inhabitant in each participating nation is introduced in *M3* instead of FIFA rank. The estimation is run for the last two World Cups, which is in itself a limitation to *M3*. Then, the host variable is automatically dropped because there are only two observations. The number of players is not significant which may be interpreted as follows: soccer mass participation is not a determinant of a nation's participation to the semi-finals of the World Cup final tournament. This invalidates for soccer the pyramidal view of sport where the larger the pyramid base of mass participation, the higher performance in international contests. On the other hand, population is significant, the semi-final history variable is even more significant (at 5%) than in previous specifications while GDP per capita and squared become significant at 10%. However, regional variables, Europe and South America, are not significant because only two World Cups are kept: in 2006, no South American team has reached the semi-finals whereas in 2002 one semi-finalist was neither European nor South American (South Korea).

Finally, a control for endogeneity between the dependent variable and one explanatory variable, the semi-final history, is required. The latter is influenced by each new World Cup results, though in the long run these results have a decreasing marginal effect on our cumulative variable. Thus, the semi-final history is used as an endogenous regressor and all other variables taken on board in *M1* as instruments. First, the semi-final history variable is regressed on population, population squared, GDP per capita and squared, hosting the Cup and regional variables, and then the

relationship between the dependent variable (making for the semi-finals) and the endogenous 'semi-final history' regressor is studied.

Insert Table 6 about here

Table 6 shows that all the instrumental variables are explanatory of the semi-final history except the host dummy. It is logical since the semi-final history variable is a cumulative percentage over 18 Cups whereas a country has been hosting the Cup only once or twice[15]. Now the model is quite consistent and close to the Olympics medal model since not only population and regional variables but also GDP per capita are significant determinants of FIFA World Cup outcome. A clear specificity is that hosting the soccer World Cup is not a comparable advantage to the one of hosting Summer Olympics. However, such reality has been blurred for a long time by the World Cup being always located either in Europe or South America until 1990. Since then, the number of exceptions has increased with one location in North America (1994), Asia (2002) and Africa (2010).

7. The prediction for the 2010 FIFA World Cup in South Africa: still so hard!

The model estimated with instrumental variables as well as *M1* specification are now used to forecast the 2010 FIFA World Cup semi-finalists, taking into account the data for population and GDP in 2008, and the cumulative semi-final history variable up to 2006. The prediction is exhibited in Table 7.

Insert Table 7 about here

The four teams with the highest probability to make for the semi-finals in South Africa are the same with both *M1* and the model with instrumental variables. If one interprets the two highest ranks (probabilities) as the most probable finalists, the former predicts Germany playing Italy in the finals while the latter forecasts Germany

playing Brazil. France is ranked fourth in both cases. Compared to FIFA ranking published in May 2010, these results are strikingly different: the first four FIFA-ranked teams are Brazil and Spain (potential finalists), then Portugal and the Netherlands. Brazil is the most widely admitted semi-finalist whatever the methodology used for prediction. If one goes as far as interpreting these rankings as a probability to participate to the 1/8th finals, there is a good chance that Argentina, Brazil, Chile, England, France, Germany, Greece, Italy, the Netherlands, Portugal, Serbia, Spain, and Uruguay would qualify for the second stage of the 2010 soccer World Cup final tournament. Since the two models encompass a host country effect, both predict South Africa qualifying for the second stage of the final tournament contrarily to this nation's FIFA ranking (83rd in May 2010). Of course, those fourteen teams[16] which are not mentioned in Table 7 would be big surprises if qualified for the semi-finals. None of them has made it!

Actually, the four semi-finalists of the 2010 World Cup have been: 1/ Spain, 2/ Netherlands, 3/ Germany, 4/ Uruguay. Thus our model did not perform with the soccer World Cup as well as it did with Olympic medals since it correctly predicted only one semi-finalist (Germany). Nobody (see below) expected Uruguay to qualify for the semi-finals while it is the fifth best probability (behind France) to qualify in our model prediction.

The banking business has recently started using predictions of the soccer World Cup outcome as an appealing factor to investors with integrating these predictions in the promotion of financial products. Consequently, some banks' economists have elaborated on prediction models that can be compared – including their results – with our model. Goldman Sachs, J.P. Morgan and UBS (*Union des Banques Suisses*) have produced a prognosis about the semi-finalists of the 2010 soccer World Cup.

Goldman Sachs (2010) has predicted the following semi-finalists, ranked according to their probability to qualify: 1/ Brazil, 2/ Spain, 3/ Germany, 4/ England - two correct out of four – with a methodology primarily based on bookmakers' odds (Ladbrokes.com) as of May 4, 2010 and partly on simulating the outcome of each qualification group and then of each of the hypothetical resulting match during the knock out stage. However some guesstimates interfere as: "from Group A, France would seem the strongest, but Mexico looks dangerous, Uruguay is a bit of an unknown, and then there are the hosts, South Africa … This could be quite a tricky group for the ageing (and some – especially Irish observers – might say undeserving!) French. I am going to assume that Mexico wins and South Africa is runner-up". Wrong anyway, but from a methodological point of view this sounds hardly more than a toss-up!

The study by J.P. Morgan (2010) adapts its QUANT scoring model (used to identify long/short trading opportunities in financial markets) to forecasting the soccer World Cup outcome by combining several footballistic variables. This scoring model delivers the following ranking: 1/ Brazil, 2/ Spain, 3/ England, 4/ Netherlands. Then, the calculated scores (for all teams) are used – excluding any tied game – together with a "penalty shoot out" metric to decide which country will win each of the 64 fixtures; the calculation comes out with a England-Spain finals won by England due to a better penalty shoot out index. This model confines itself to FIFA World Cup variables but does not perform better, predicting only two of the actual semi-finalists.

UBS (2010) approach states from the very beginning that "socioeconomic factors like population size or GDP growth have no explanatory power when it comes to forecasting the performance of a specific team" and that "at every World Cup there is at least one surprise participant in the semi-finals". UBS model takes on board: a

team's past performance in the World Cup; whether or not a team is a host nation; an objective quantitative measure that assesses the strength of each team three months before the start of the World Cup. The last variable is calculated by using the Elo ratings developed to measure and rank the strength of chess players; it is assumed to be better than FIFA ranking because it takes into account not only the number of a team's wins, losses and draws, but also the specific circumstances under which those events occurred. Brazil is predicted to have the highest probability to win the 2010 World Cup (Spain has only the 7th best probability). The best probabilities to make for the semi-finals are: 1/ Brazil, 2/ Germany, 3/ Netherlands, 4/ Italy. Still 50% correct predictions are found – which also means 50% wrong.

8. Sport surprising outcomes and its metrics: an avenue for further research

Unexpected or surprising outcomes of a sport game or contest have not really been analysed so far. The first point to clarify is the difference between the concept of outcome uncertainty and a surprising outcome. On the one hand, the uncertainty of outcome basically is an ex ante concept – it results from the equality or closeness of sporting forces which are going to be opposed in a game or a sport contest – while a surprising outcome is necessarily an ex post notion: the actual outcome has appeared surprising compared to some ex ante expectation or prediction or standing. A surprising outcome is, to some extent, the opposite of outcome uncertainty which is deeply rooted in outcome unpredictability. The latter is very high when two teams are so close in terms of sporting forces that it is impossible to predict the game outcome (or all teams are so close that the league's final ranking cannot be predicted). A surprising outcome is quite the opposite insofar as it occurs when a sporting outcome

is rather predictable but happens to be different from the prediction. This happens when opponents in a game (contest) have clearly uneven sporting forces, and the underdog wins the favourite, for instance a low FIFA-ranked national team defeats a high FIFA-ranked nation. In a nutshell, a surprising sport outcome may be defined as the *ex post invalidation of an ex ante rather high outcome certainty (predictability)* whereas outcome uncertainty is the ex ante unpredictability of an ex post actual outcome.

Many metrics may be conceived for measuring the occurrence of a surprising sporting outcome. This is an avenue for further research and as a first step macro- and micro-assessments of a surprising sporting result can be suggested[17]. With the aforementioned FIFA World Cup prediction model, a macro-surprise would occur when a team had not made it for the semi-finals while the model was predicting its qualification – and symmetrically when an unpredicted team qualified for the semi-finals. As to this model, Spain and the Netherlands qualification (to a lesser extent Uruguay qualification) were surprising as well as Brazil, Italy and France not making it for the semi-finals. To obtain an overall metrics, suffice it to put that when a team is higher (lower) ranked by the model than the actual ranking of the final tournament, one witness a surprising sport macro-outcome. The surprise magnitude can be assessed by the rank difference between the model's prediction and the actual outcome (Table 8). In a same way, it would be possible to define macro-surprises comparing FIFA ranking and FIFA World Cup outcome or comparing the latter with banks' predictions.

Insert Table 8 about there

With all predictions, three big surprises emerged: Ghana, Paraguay and Japan making it for the 1/8th finals. Uruguay also is a rather big surprise, since its

qualification for the 1/8th finals and even the ¼ finals were only predicted with our model. To some extent England's ranking due to a severe loss (1-4) against Germany in the 1/8th finals was also surprising. Notice that Goldman Sachs did not find any big surprise while the variance between its predictions and achieved outcomes is higher than with our model. The latter detects two big surprises: Uruguay qualifying for the ¼ finals and the failure of England's team. JP Morgan's predictions have the highest variance with the achieved outcomes whereas it is confirmed that FIFA ranking is not a good predictor either.

However, differences in the exact meaning of measured surprises must be underlined. A comparison between the actual FIFA World Cup outcome and our model's predictions exhibits sporting surprises with regards to nations' economic development and their past performances in the World Cup. With JP Morgan's prediction and FIFA ranking, strictly speaking 'footballistic' surprises are pointed at: the actual outcome is surprising compared to exclusively 'footballistic' variables. Goldman Sachs' prediction shows how much bookmakers' odds before the World Cup were distant from the achieved outcomes. Betting, in accordance with Goldman Sachs, on Germany as the Cup winner would have resulted in a gambler's monetary loss while betting on Spain as one of the finalists would have yielded some return.

A micro-metrics of surprising sport outcomes may be based on a weaker team (underdog) winning a stronger team (favourite). If one lower FIFA-ranked team won a higher FIFA-ranked team, this would be a micro-surprise. With our model, Ghana-USA (2-1) in the 1/8th finals was surprising as well as Netherlands-Brazil (2-1) in the ¼ finals and Spain-Germany (1-0) in the semi-finals. With FIFA ranking and Goldman Sachs' prediction, the micro-surprises were only Ghana-USA (2-1) and Netherlands-Brazil (2-1). On the other hand, JP Morgan's prediction was surprised by Ghana-USA

(2-1) and Germany-England (4-1) in the 1/8th finals and Netherlands-Brazil (2-1) and Germany-Argentina (4-0) in the ¼ finals. JP Morgan's prediction definitely is the furthest from actual outcomes when looking at both micro- and macro-surprises.

The notion of surprising sporting micro-outcome may be refined with the historical variable of our model, checking whether in the 2010 World Cup final tournament a nation which never made it for the semi-finals has been able to win a nation which already qualified for the semi-finals at least once since 1930. With such a criterion, following outcomes are surprising: Ghana-USA (2-1) in the 1/8th finals and during the round robin stage Mexico-France (2-0), South Africa-France (2-1), Serbia-Germany (1-0), Slovakia-Italy (3-2), and Switzerland-Spain (1-0). Two teams did not survive to their surprising losses in the qualification groups (France and Italy) whereas the two others even made it for the semi-finals (Germany and Spain).

Conclusion

Since our modelled prediction had been able to correctly detect 70% of actual medal winners at the Beijing Games, a nation's size (population) and level of economic development, once completed with a few dummies, are good predictors of medal wins. The latter can be taken as a relevant index for comparing economic development across nations in addition to other economic and social indexes. A same model does not perform that well with predicting the 2010 FIFA World Cup semi-finalists. Soccer World Cup outcomes are in no way an acceptable index of economic development. The host country effect (home advantage) is less significant in soccer World Cup than in Summer Olympics. However, any economic prediction of sporting performance must be taken with a pinch of salt. This is namely due to a

number of surprising sporting outcomes. Elaborating on a metrics to quantify them should be a promising avenue for further research.

References:

Andreff M., W. Andreff & S. Poupaux (2008), Les déterminants économiques de la performance olympique: Prévision des médailles qui seront gagnées aux Jeux de Pékin, *Revue d'Economie Politique*, 118 (2), 135-69.

Andreff W. (2001), The Correlation between Economic Underdevelopment and Sport, *European Sport Management Quarterly*, 1 (4), 251-79.

Arellano M. & S. Bond (1991), Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, 277-97.

Ball D. (1972), Olympic Games Competition: Structural Correlates of National Success, *International Journal of Comparative Sociology*, 13, 186-200.

Bernard A.B. & M.R. Busse (2004), Who Wins the Olympic Games: Economic Resources and Medal Totals, *Review of Economics and Statistics*, 86 (1), 413-17.

Clarke S.R. (2000), Home Advantage in the Olympic Games, in G. Cohen & T. Langtry, eds., *Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport, Conference proceedings*, Sydney: University of Technology Sydney, 43-51.

Coupé T. (2007), Incentives and Bonuses – The Case of the 2006 World Cup, *Kyklos*, 60 (3), 349-358.

Goldman Sachs (2010), *The World Cup and Economics 2010*, Goldman Sachs Global Economics, Commodities and Strategy Research, May.

Grimes A.R., W.J. Kelly & P.H. Rubin (1974), A Socioeconomic Model of National Olympic Performance, *Social Science Quarterly*, 55, 777-82.

Groot L. (2008), *Economics, Uncertainty and European Football. Trends in Competitive Balance*, Cheltenham: Edward Elgar.

Hill D. (2009), How Gambling Corruptors Fix Football Matches, *European Sport Management Quarterly*, 9 (4), 411-32.

Hoffmann R., L.Chew Ging & B. Ramasamy (2002a), Public Policy and Olympic Success, *Applied Economic Letters*, 9, 545-48.

Hoffmann R., L.Chew Ging & B. Ramasamy (2002b), The Socio-Economic Determinants of International Soccer Performance, *Journal of Applied Economics*, 5, 253-72.

Houston R.G. Jr & D.P. Wilson (2002), Income, Leisure and Proficiency: An Economic Study of Football Performance, *Applied Economic Letters*, 9, 939-43.

Johnson D. & A. Ali (2004), A Tale of Two Seasons: Participation and Medal Counts at the Summer and Winter Olympic Games, *Social Science Quarterly*, 85 (4), 974-93.

J.P. Morgan (2010), *England to Win the World Cup! A Quantitative Guide to the 2010 World Cup*, J.P. Morgan Europe Equity Research, May.

Leeds M. & E. Marikova Leeds (2009), International Soccer Success and National Institutions, *Journal of Sports Economics*, 10 (4), 369-90.

Levine N. (1974), Why Do Countries Win Olympic Medals? Some Structural Correlates of Olympic Games Success: 1972, *Sociology and Social Research*, 58, 353-60.

Macmillan P. & I. Smith (2007), Explaining International Soccer Rankings, *Journal of Sports Economics*, 8 (2), 202-13.

Maennig W. & Wellebrock C., (2008), Sozioökonomische Schätzungen olympischer Medaillen-gewinne. Analyse-, Prognose- und Benchmarkmöglichkeiten. *Sportwissenschaft* 2, 131-48.

Monks J. & J. Husch (2009), The Impact of Seeding, Home Continent, and Hosting on FIFA World Cup Results, *Journal of Sports Economics*, 10 (4), 391-408.

Nevill A., G. Atkinson, M. Hughes & S. Cooper (2002), Statistical Methods for Analyzing Discrete and Categorial Data Recorded in Performance Analysis, *Journal of Sports Sciences*, 20 (10), 829-44.

Novikov A.D. & A.M. Maximenko (1972), The Influence of Selected Socio-economic Factors on the Levels of Sports Achievements in the Various Countries, *International Review of Sport Sociology*, 7, 27-44.

Paul S. & R. Mitra (2008), How Predictable Are the FIFA Worldcup Football Outcomes? An Empirical Analysis, *Applied Economic Letters*, 15, 1171-76.

Poupaux S. & W. Andreff (2007), The Institutional Dimension of the Sports Economy in Transition Countries, in M.M. Parent & T. Slack, eds., *International Perspectives on the Management of Sport*, Amsterdam: Elsevier, 99-124.

Rathke A. & U. Woitek (2008), Economics and the Summer Olympics: An Efficiency Analysis, *Journal of Sports Economics*, 9 (5), 520-37.

Torgler B. (2004), The Economics of the FIFA Football Worldcup, *Kyklos*, 57 (2), 287-300.

UBS (2010), *UBS investor's guide. Special edition: 2010 World Cup in South Africa*, UBS Wealth Management Research, April.

Yamamura E. (2009), Technology Transfer and Convergence of Performance: An Economic Study of FIFA Football Ranking, *Applied Economics Letters*, 16, 261-266.

[1] Bernard and Busse use the percentage of medal wins by each country *i* for $M_{i,t}$ instead. Our regressions are calculated with both the absolute number of medals (Table 3) and the percentage of medals per country, and the results are not significantly different.

[2] Hoffmann *et al.* (2002a) consider that an important determinant of Olympic successes lies in the degree to which sporting activities are embedded in a nation's culture. The proxy used to capture such determinant is the total number of times a country has hosted Summer Olympics from 1946 to 1998.

[3] A discussant has suggested to test in a first stage a "winning versus not winning a medal" hypothesis and then estimate, in a second stage, the number of medal wins (when > 0). Here we assume that winning zero medal or winning 1, 2, …, n medals results from the same procedure and must be estimated with the same explanatory variables.

[4] A test of maximum likelihood shows that the rho coefficient is significant (*Pr = 0.00*).

[5] Our data panel is not balanced since the number of participating countries has increased between 1976 and 2004, namely due to the break up of the former Soviet Union, former Yugoslavia and former Czechoslovakia (+ 20 countries in the world), only partly compensated by the re-unification of Germany and Yemen (- 2 countries).

[6] A test of maximum likelihood shows that the rho coefficient is not significant (*Pr = 0.26*) which allows to opt for a pooling estimation.

[7] Result for any other country is available on request addressed to the authors.

[8] This issue is discussed in depth in Andreff *et al.* (2008) explaining why we had not been able to integrate doping among independent variables despite that we wished to do so.

[9] Some Jamaican sprint finalists have been controlled positive in doping tests during the weeks after the Beijing Games, which may be another explanation.

[10] The role of referees is neglected here for two reasons: an imperfect referee is a source of competitive unbalance as demonstrated in Groot (2007), and a corrupt referee paves the way for another kind of study about corruption in soccer. We make the (rather naive) assumption that there is no match fixing and no rigged games even though it is definitely a simplifying assumption in current international soccer (Hill, 2009).

[11] All the results are obviously plagued with endogeneity since the final standing is correlated with ex ante ranking and top seeding is determined by ex ante ranking. No methodology is implemented to clean or circumvent it.

[12] Exceptions are team sports and some other sports such as tennis and table tennis.

[13] Here participating countries refer to those qualified for the soccer World Cup final tournament. Our model does not attempt to estimate the determinants of this qualification.

[14] The calculation formula of FIFA ranking encompasses, among other, a weighted average of the team's three previous FIFA World Cup results.

[15] The FIFA World Cup final tournament has been hosted twice in France (1938, 1998), Germany (1974, 2006), Italy (1934, 1990) and Mexico (1970, 1986).

[16] Algeria, Australia, Cameroon, Denmark, Ghana, Honduras, Ivory Coast, Japan, New Zealand, Mexico, Nigeria, Paraguay, RDP (North) Korea, and Switzerland.

[17] Micro here means at a one game level, and not an aggregated result as the final ranking of the World Cup final tournament – which defines a macro-surprise (coming out from 64 fixtures).