Forecasting National Medal Totals at the Summer Olympic Games Reconsidered

Nicolas Scelles, Wladimir Andreff, Liliane Bonnal, Madeleine Andreff, Pascal Favard

Dr Nicolas Scelles, Manchester Metropolitan University (corresponding author)
Manchester Metropolitan University Business School,
All Saints Campus, Oxford Road
Manchester, M15 6BH
UK
N.Scelles@mmu.ac.uk

Prof Wladimir Andreff, University Paris 1 Panthéon Sorbonne
Centre d'Economie de la Sorbonne, Maison des Sciences Economiques
106-112 bd de l'Hôpital
75642 Paris Cedex 13
France
andreff@club-internet.fr

Prof Liliane Bonnal, University of Poitiers
Université de Poitiers, UFR Sciences Economiques, Bâtiment A1
2 rue Jean Carbonnier
TSA 81100 – 86073 Poitiers Cedex 9
France
liliane.bonnal@univ-poitiers.fr

Mrs Madeleine Andreff, University Paris-Est Marne-la-Vallée
UFR de Sciences Economiques et de Gestion, Université Paris-Est Marne-la-Vallée
Rue Galilée, Bâtiment Bois de l'Etang
77454 Champs-sur-Marne Cedex 02
France
madandreff@gmail.com

Prof Pascal Favard, University of Tours
UFR de Droit, Economie et Sciences Sociales, Université de Tours
50 avenue Jean Portalis, Les deux Lions
BP 0607 – 37206 Tours Cedex 03
France
pascal.favard@univ-tours.fr

Funding
This work was funded by the French Observatory of the Sports Economy, Ministry for Sports.
Conflict of interest
None
Acknowledgements
The authors would like to thank the French Observatory of the Sports Economy, French Ministry for Sports, for having funded this research. They also want to thank the participants of the 11th Conference of the European Sport Economics Association in Gijon, Spain, for their feedback on the presentation of the paper. In particular, they thank Prof David Forrest and Dr Juan de Dios Tena for their feedback on the full paper. The authors also want to thank the editors and the referee for their very helpful comments. Errors are their own.

Abstract:

Objective. This paper aims at explaining national medal totals at the 1992-2016 Summer Olympic Games (n = 1289 observations) and forecasting them in 2016 (based on 1992-2012 data) and 2020 with a set of variables similar to previous studies, as well as a regional (sub-continents) variable not tested previously in the literature in English. *Method.* Econometric testing not only resorts to a Tobit model as usual but also to a Hurdle model. *Results.* Most variables have a significant impact on national team medal totals; it appears to be negative for most regions other than North America except Western Europe and Oceania (not significant). Then, two models (Tobit and Hurdle) are implemented to forecast national medal totals at the 2016 and 2020 Summer Olympics. *Conclusion.* Both models are complementary for the 2016 forecast. The 2020 forecast is consistent with Olympic Medals Prediction (2020), although some striking differences are found.

Keywords: Tobit regression, Hurdle regression, sport performance.

The 2020 Summer Olympic Games will take place in Tokyo. Consistent with what happens before each Olympics edition, predictions about the national team medal totals have been made, based on the latest observed sporting results (Olympic Medals Predictions, 2019). The problem with such predictions is that they do not inform about the socioeconomic, political and sporting determinants explaining why such sporting results are supposed to come out. In the academic literature, a number of research works have attempted to explain medal win distribution at Summer Olympics with models that encapsulate the aforementioned types of variables. In a similar vein, this paper aims at explaining previous national team medal totals at the 1992-2016 Summer Olympic Games (n = 1289 observations) with a similar set of variables though including the test of a regional variable which has not been taken on board in the literature in English so far, although Andreff, Andreff and Poupaux (2008) tested it in an article in French aiming at forecasting medal totals at the 2008 Beijing Olympics. Another objective is to work out econometric testing not only resorting to a Tobit model as usual but also to a Hurdle model. Two models (Tobit and Hurdle) are then implemented in such a way as to forecast national team medals totals at the 2016 (based on the results from 1992-2012) and 2020 (based on the results from 1992-2016) Summer Olympics.

The article reads as follows: first, a literature review enables to identify potential explanatory variables; second, a new model, and its variants, is presented; third, the results of our explanatory models are exhibited for the 1992-2016 period; fourth, derived forecasting models are tested over the same period of time; fifth, forecasts for the 2016 Summer Olympic Games are provided; sixth, estimated forecasts for the 2020 Summer Olympic Games are exhibited and then compared with estimates published in Olympic Medals Predictions as of January 28, 2020. The last section concludes.

Literature Review

Explaining Summer Olympics medal win distribution and, consequently, national medal totals is not a brand new train of thought. This kind of exercise started as early as in the 1970s though an important step forward was achieved in 2004. That year, Bernard and Busse (2004), comparing the different econometric methodologies, came up with the conclusion that a Tobit model always delivers better results. Then it became standard to estimate an explanatory model of medal wins distribution with a Tobit (e.g. Andreff *et al.*, 2008; Forrest, Sanz & Tena, 2010; and so on) and, since Bernard and Busse geared their article towards prediction as well, the Tobit regression turned out to be the hard core methodology in forecasting national medals totals.

Bernard and Busse (2004), working with panel data on the 1960–96 Summer Games, first estimated a model which explains a nation's share in the total number of medals. Probit and Tobit regressions were used. The hypothesis that medal winning should be proportional to population was econometrically rejected. Interestingly, per capita income and population were found to have very similar and significant effects at the margin on the production of Olympic medals. This suggests that total GDP is the best predictor of national Olympic performance. The model was then used to predict the number of medals won by Australia in 2000, and the result was only slightly different from the observed total. Bernard and Busse concluded that forced mobilisation of resources by governments can also play a role in medal total - an argument that probably applies in retrospect to past Soviet and Eastern European Olympic performances too.

Fully in tune with Bernard and Busse, Andreff *et al.* (2008) modelling took on board GDP per capita, population, a host effect and a political regime variable delineating more precise sub-samples among the post-communist economies than in Bernard-Busse's article. An additional regional variable was supposed to capture different sports specialisation in different regions (sub-continents) of the world economy, namely *NAM* (North America), *AFN* (North Africa), *AFS* (Sub-Saharan Africa), *LSA* (Latin and South America), *EAST* (Eastern Europe), *WEU* (Western Europe), *OCE* (Oceania), *MNE* (Middle East), and *ASI* (Asia). The dependent variable, in contrast with Bernard-Busse, was national medal totals rather than a country share (percentage) in the total medal distribution. It appeared that adding a variable standing for the number of medals won by each country at the previous Olympics (in *t-4* for the Olympics in *t*) markedly improved the censored Tobit econometric results, as already shown by Bernard-Busse; the underlying rationale is that, to a non-negligible extent, past Olympic successes are predictors of current Olympic performances. This was a useful lesson for those running models with a view to forecasting forthcoming national medals totals. Notice that GDP per capita and population are four year lagged (values taken in *t-4*) with the underlying assumption that a given span of time is required to prepare an Olympic team, here assessed to be four years; put otherwise, as soon as the *t-4* Olympics are over, each national team starts preparing for the *t* Olympics. By the same token, some inertia is introduced this way into the model which may avoid explosive variations when it is used for forecasting. Interestingly, compared to *WEU*, the regional variable unveils a significant positive impact for *AFS*, *NAM* and *OCE*, no significant impact for *LSA* and a significant negative impact for *AFN*, *ASI*, *EAST* and *MNE*. Despite this variable being significant, it has not been used since Andreff *et al.* (2008), maybe because this article is in French and, as such, not taken into account in the literature reviews conducted by authors focusing on papers in English.

Andreff *et al.* (2008) published their article prior to the 2008 Beijing Olympics. Andreff (2009) compared their forecasts with actual medals, finding that Andreff *et al.* (2008) predicted correctly 70% of the medals with a 95% interval confidence and even 88% with a two medals error margin. Andreff (2009) identified doping as the explanation for those countries for which forecasts were not accurate.

Forrest *et al.* (2010) adapted the Bernard-Busse model to include two new covariates, namely the level of public expenditure on recreational, cultural and religious affairs (including sport) in each country provided by the United Nations (UN) and whether future hosts of the Games have such a great incentive to raise their performance standards that this is already reflected in their achievements in the current Olympiad. Both variables have a significant positive impact on the shares of medals for the 1992 to 2004 Olympics. The authors then attempted to forecast national team medal totals at the 2008 Beijing Olympics. To do so, they made subjective, judgemental adjustments, for example that the extra medals attributable to the old way of doing things for the post-communist economies will fade away over time, which is confirmed ‘objectively’ by Forrest, McHale, Sanz and Tena (2015, 2017) and Noland and Stahler (2016, 2017).

Vagenas and Vlachokyriakou (2012) looked at the predictors of medal totals at the 2004 Olympics. They tested two new variables, namely the impact of having hosted the Games four years earlier and the number of participant athletes per country. For both variables, they found a significant positive impact. Also introducing a new variable, Vagenas and Palaiothodorou (2019) exhibited empirical evidence contrary to the hypothesis of climatic impact on Olympic performance, in particular no superiority of temperate climate nations shows up from a Tobit testing on six Summer Games (1996-2016). Leeds and Leeds (2012), Trivedi and Zimmer (2014) and Lowen, Deaner and Schmidt (2016), looking at the impact of gender (for the first two) or gender inequalities (for the latter), did not find any significant result.

Blais-Morisset, Boucher and Fortin (2017) attempted to explain a nation medals total for the 1992 to 2012 Olympics. The chosen dependent variable is discrete, and drives the authors to estimate a Poisson model and then a negative binomial model, including a Zinb (zero inflated negative binomial) model specification rather than a Tobit as in most previous studies. Similar to Forrest *et al.* (2010), the authors tested the impact of the level of public expenditure on recreational, cultural and religious affairs. They found that it is a better indicator of Olympic performances than GDP per capita. The authors interpret their result as public investment in sports being a better targeted governmental policy tool in view to gaining a nation's successes at the Olympics. Extremely topical and interesting, such result is to be taken with a pinch of salt due to a serious limitation. Indeed, it has been found with a sample of 53 nations that is roughly one quarter of all participating nations in the last Olympics.

Compared to the aforementioned studies, Celik and Gius (2014), studying the 1996-2008 Olympics, used a different dependent variable: instead of the number (or the share in total) of national medal totals at the end of the Games, they subtracted those medals stripped off from athletes ex post disqualified for doping. Otherwise, their model was basic with population, GDP per capita, host effect and the number of medals awarded at the previous Games, the latter improving the forecast of national medal totals once cleaned from disqualifications.

Otamendi and Doncel (2018) raised the issue of whether the medal win distribution is better anticipated by forecasting models or by sports experts who have a deep knowledge of the different Olympic sport disciplines. They compared five expert predictions published in the press with three forecasting models respectively used for the 2010 Vancouver Winter Olympics (Otamendi & Doncel, 2014a), the 2012 London Summer Olympics (Otamendi & Doncel, 2014b) and the 2014 Sochi Winter Games (Andreff, 2013). Relying on indicators to test the performance of a forecast such as a ratio of exactly predicted results, Pearson, Kendall and Spearman correlations adjusting the forecast of ex ante statistical distribution to the ex post observed one, the authors concluded that sports experts' predictions are more accurate as regard the detailed medal distribution within a given sport discipline while econometric models perform better when it comes to medal wins distribution across participating nations. Otamendi and Doncel (2018)'s final comment suggests that expert forecasts are more to be used by sport punters whereas econometric forecasts are more useful for designing public sport policies. The latter is the outlook of the modelling adopted below.

Data and Methodology

What is intended here is to compare from a forecasting perspective the results of estimating a Tobit and a Hurdle model, in panel with random effects for both. Data have been gathered for all Games from Barcelona 1992 up to Rio de Janeiro 2016 (n = 1289 observations).

Variables

First of all, the dependent variable $Mapdisq_{i,t}$ is, for nation i in year t, a corrected number of medal wins which may not be equal to the actual number of medals won and publicised right after ending the Games. $Mapdisq_{i,t}$ is a national medals total **after** deducing all ex post medals lost due to (often doping) disqualifications of nation i's athletes[1]. Data are from https://en.wikipedia.org/wiki/Summer_Olympic_Games, the Summer Olympics Wikipedia English site which links to web pages of different Games where tables are found regarding medals totals, medallists' disqualifications, and medals' reallocations; references to IOC official data are reported so that double-checking can be done. It is worth noting that a better assessment and accounting of the doping impact on Olympic performances would require information about the number of all doped (including non-detected) athletes which would enable to use doping as an explanatory variable instead of using it as an alleviation of the dependent variable. Such information has no chance to be unveiled in any foreseeable future (Andreff, 2019). Therefore, doping remains a non-observable – and widely unobserved[2] – variable in view to explaining and forecasting national medal totals so far.

Turning now to other variables, six basic explanatory variables significant in Bernard and Busse (2004) and Andreff *et al.* (2008) works are kept as our model's hard core:

1/ $N_{i,t-4}$ stands for population in participating country i four years earlier than year t Olympics. Data are collected from the World Bank: https://data.worldbank.org/indicator/SP.POP.TOTL, and the variable logarithm is used in estimating our model variants.

2/ $(Y/N)_{i,t-4}$ stands for gross domestic product (GDP) per inhabitant in nation i four years earlier than year *t* Olympics and is assumed to capture its level of economic development, differentiating rich/developed and poor/developing countries.

These first two variables are taken four years earlier under the assumption that nation *i* needs to mobilise economic and demographic resources four years in advance to prepare its Olympic team and have it ready for the year *t* Olympics. In the background, the rationale is that human and economic resources need to be available from the starting point of the national Olympic team's preparation for the next Games that we assume to start up right after the end of previous Games, *i.e.* four years earlier.

Data are constant purchasing power parity Gross Domestic Product, in 2011 million US international dollars and data is collected from the CEPII database called CHELEM open on the DBnomics site: https://db.nomics.world/CEPII/CHELEM-GDP, except for Puerto Rico (absent in the database). For the latter country, constant PPP GDP has been found in the World Bank database.

3/ $Host_{i,t}$ is a dummy variable supposed to capture a host country effect on medal wins and is equal to 1 for host countries and equal to 0 for other participating nations.

4/ $Political\ Regime_{p,i}$ is a dummy that differentiates among participating nations between former socialist centrally-planned economies, *i.e.* Central Eastern European countries (*CEEC*), that have joined the European Union, then all other (post-)communist economies (*POSTCOM*), and capitalist market economies (*CAPME*) which all other countries in the world are assumed to be. However, in most recent studies (Forrest *et al.*, 2015, 2017; Noland & Stahler, 2016, 2017) post-communist transition economies did benefit much less from their outlier[3] situation than at the dawn of transition period or before it, when Soviet-style sports were very much supported by the state to win medals. Consequently, the $Political\ Regime_{p,i}$ variable classifies all participating nations into three country groups:

CEEC: 11 post-communist nations which joined the EU (Bulgaria, Croatia, the Czech

Republic[4], Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, and Slovenia).

POSTCOM: 23 other (post-)communist nations which are not EU members (Albania, Armenia, Azerbaijan, Belarus, Bosnia-Herzegovina, China, Cuba, Georgia, Kazakhstan, Kosovo, Kyrgyzstan, Laos, Macedonia, Moldova, Mongolia, Montenegro, People's Republic of (North) Korea, Russia, Serbia[5], Tajikistan, Ukraine Uzbekistan, and Vietnam).

CAPME: capitalist market economies, without differentiation, assuming that all other participating nations are such economies; this country group is taken as the reference.

5/ $Regions_{r,i}$ is a dummy which classifies each nation i into one of the nine following country classes : *NAM* (North America), *AFN* (North Africa), *AFS* (Sub-Saharan Africa), *LSA* (Latin and South America), *EAST* (Eastern Europe), *WEU* (Western Europe), *OCE* (Oceania), *MNE* (Middle East), and *ASI* (Asia). Following up Andreff *et al.* (2008), this variable is assumed to be a proxy for nations' cultural and regional specialisation in some given sports disciplines, common to several countries in a same region in the world[6].

6/ Medal totals four years earlier $M_{i,t-4}$ is the actual number of medals won by nation i at previous Games net of ex post disqualifications. This variable is taken on board to make our model ergodic and because it improves more than slightly medal win forecasts (Bernard & Busse, 2004; Celik & Gius, 2014); it is introduced only in forecasting variants of the model.

Beyond these six variables, three other variables have been tested to check whether their explanatory power makes it worth including them in our model: the number of participating athletes per national team; hosting the Games four years later; and having hosted the Games four years earlier.

1/ $NA_{i,t}$ stands for the number of participating athletes in each national team i, the rationale being that countries fielding more athletes are more likely to win medals. Data are drawn from the Wikipedia site:

https://en.wikipedia.org/wiki/2016_Summer_Olympics#Participating_National_Olympic_Co mmittees for each Olympic Games from 1992 to 2016. It is first tested as a continuous variable. Then it is tested as a discrete variable $RNA_{i,d,t}$ which splits the number of participating athletes into four classes (from 0 to 9 athletes, from 10 to 49 athletes, from 50 to 149 athletes, and 150 athletes and over), for two reasons. For the one, from an analytical standpoint, marginal return to the number of participating athletes may not be constant.

Making the variable discrete enables dropping a constant return assumption. On the other hand, such discrete variable enables having some information about the potential number of participating athletes that can be used when forecasting national medals totals at the next Games without knowing ex ante the exact number that each national team will actually field.

Notice that athlete selection in the host country's Olympic team obeys specific criteria (lower sporting performance requirements) with ensuing consequence that a nation fields a bigger number of athletes when it hosts the Games than otherwise. Obviously the two variables - host country and number of participating athletes - are linked. Three different models are required to disentangle them, taking on board respectively: 1/ the host country effect alone (model 1 below); 2/ only the number of participating athletes as a continuous variable (model 2 below); and 3/ the two variables together while considering the four athlete classes – a discrete variable (model 3 below, used for forecasting).

2/ A second variable $Host\ in\ 4\ years_{i,t}$ stands for the impact on a nation *i*'s Olympic performance of its knowledge that it will be hosting the next Games four years later. The underlying assumption is that, being the next organising host country, this nation's athletes will start up training and preparing themselves in advance with the objective of achieving very high level Olympic performances when they will benefit from the host effect. Usually, the Games are awarded to a city/country about seven years in advance (*t-7*), thus an early preparation of the Olympic team may be beneficial in terms of medal wins as early as in the next Games in *t-4*. Such effect was mentioned, for example, when explaining why the British team was so much successful (47 medals) at the 2008 Beijing Olympics. Maennig and Wellbrock (2008) tested a so-called “Great Britain will host the 2012 Olympics” variable as significantly positive.

3/ A third and last variable *Host 4 years ago*$_{i,t}$ stands for having hosted the Games four years earlier, the rationale being that the investment made in view to winning many medals during the Games hosted in *t* should still affect positively the host country’s Olympic performance four years later in *t+4*. The intuition is as follows: intensively preparing and training athletes to win more medals when a nation is hosting the Games in *t* may have lasting beneficial effects up to the next Games when the nation is no longer the host country. Thus the *Host 4 years ago*$_{i,t}$ variable is equal to 1 for a nation *i* when it had been the previous Games organising country. For instance, taking Great Britain as an example, hosting the 2012 London Games translates in our models into *Host in 4 years*$_{i,t}$ = 1 for 2008, *Host* = 1 for 2012 and *Host 4 years ago*$_{i,t}$ = 1 for 2016.

Tobit and Hurdle Modelling

As mentioned previously, we have estimated both Tobit and Hurdle models. The use of a Tobit model is justified by the large mass points at zero medal (Bernard & Busse, 2004; Forrest *et al.*, 2017). As noted by Forrest *et al.* (2017), the data are therefore treated as subject to censoring, which is intuitive because some countries come closer than others to winning a medal, for example they win some fourth-places, yet the performances of all of them are recorded as zero. In their article, these authors choose to use a Tobit model for three reasons. First, this facilitates comparison with Bernard and Busse (2004). Second, it is hard to think of theoretical reasons why the Tobit model would be inappropriate since it appeared to them plausible that the same mechanisms (resources) would drive both whether a country would win medals and how many it would win if it did. Third, their focus was to be on individual sports with comparisons across them based on medal shares rather than medal counts to control for the different numbers of medals available in each sport at each Games; therefore, a count model such as Poisson would make comparisons across sports not straightforward.

By contrast with Forrest *et al.* (2017), we are not interested in individual sports with comparisons across them and use medal counts. Therefore, we can also test a model that explicitly accounts for the discrete nature of the dependant variable, i.e. the number of medals (Blais-Morisset *et al.*, 2017). As suggested by Blais-Morisset *et al.* (2017), a count model such as Poisson can be used. This is in particular possible when the dependant variable takes discrete values that are quite low, as this is the case for most countries regarding the number of medals won at the Olympic Games. Poisson models have the particularity to assume that the expected value and variance of the random variable are equal. Such hypothesis is relatively constraining and not realistic in our case since there is a strong heterogeneity across countries. In order to account for this heterogeneity, a negative binomial model is considered, which generalises the Poisson model by introducing in the expected value an unobserved individual effect. Given that the number of countries winning no medal is quite important, a Zinb model could have been chosen, consistent with Blais-Morisset *et al.* (2017). However, such model assumes that the zero observations have two different origins (Hu, Pavlicova & Nunes, 2011): "structural" (e.g. a country does not take part in the Olympic Games) and "sampling" (e.g. a country takes part and scores zero medal at the Olympic Games). This model is not appropriate for our research since the focus is on countries having taken part in the Olympic Games.

Eventually a Hurdle model is estimated. Contrary to the Zinb model, it does not assume that the zero observations have two different origins (Hu *et al.*, 2011). Similar to the Tobit model, the Hurdle model accounts for the probability of winning no medal and for the number of medals; its advantages compared to the Tobit model are that it distinguishes between two equations (medal(s) or not then number of medals for countries winning at least one medal, only the second equation being released later in the results) and explicitly accounts for the discrete nature and the asymmetric distribution of the dependent variable. It remains to observe whether this translates in forecasts that are more accurate. Thus, all regressions are estimated with both Tobit and Hurdle models, tested in panel with random effects, in which $Mapdisq_{i,t}$ is the number of medals won by country *i* at the Games organised in *t*.

For the Tobit model, the general specification is:

$$Mapdisq^*_{i,t} = X_{i,t}\Theta + u_i + \epsilon_{i,t},$$

where $u_i \sim N(0,\sigma^2_u)$ and $\varepsilon_{i,t} \sim N(0,\sigma^2_\varepsilon)$, and $Mapdisq_{i,t} = \begin{cases} Mapdisq^*_{i,t} & if \quad Mapdisq^*_{i,t} > 0 \\ 0 & if \quad Mapdisq^*_{i,t} \leq 0. \end{cases}$

For the Hurdle model, the general specification for the part related to the count process[7] is:

$$\chi_{i,t} = exp\left(X_{i,t}\Theta + \tau_{i,t}u_i\right),$$

with: $Mapdisq_{i,t} \sim$ Poisson($\chi_{i,t}$); $\chi_{i,t}|u_i \sim$ Gamma(exp($g_{i,t}$)); $u_i \sim N(0,\sigma^2_u)$.

Depending on the set of explanatory variables selected, $X_{i,t}\Theta$ is defined by:

$$c + \alpha \ln N_{i,t-4} + \beta \ln\left(\frac{Y}{N}\right)_{t-4} + \gamma Host_{i,t} + \sum_p \delta_p PoliticalRegime_{p,i} + \sum_r \rho_r\, Regions_{r,i} + \lambda Host\ in\ 4\ years_{i,t} + \mu Host\ 4\ years\ ago_{i,t}\ \mathbf{(1)}$$

$$c + \alpha \ln N_{i,t-4} + \beta \ln\left(\frac{Y}{N}\right)_{t-4} + \sum_p \delta_p PoliticalRegime_{p,i} + \sum_r \rho_r\, Regions_{r,i} + \nu NA_{i,t}\ \mathbf{(2)}$$

It is worth noting that the number of participating athletes affects the impact of the three hosting variables, which loose statistical significance when taken on board together with the number of participating athletes. This explains why the three hosting variables are not included in Model (2). Table 1 presents summary descriptive statistics for the covariates included in the models (n = 1289 observations).

Table 1

Results of Explanatory Models

The results obtained with both Tobit and Hurdle models show that most variables have a significant impact on the medal totals (Table 2): the impact is positive for population and GDP per capita four years earlier, the two specific post-communist political regimes, the usual host effect, hosting the Games four years later, having hosted the Games four years earlier, the number of participating athletes; it is negative for most regions other than North America except Western Europe and Oceania (not significant).

Table 2

Results of Forecasting Models

The number of participating athletes cannot be directly used in forecasting models since the number of participants in each national Olympic squad is not known yet. However, the importance of this variable as a medal win determinant leads us to take it on board in forecasting models though in a different manner: the variable is made discrete by means of grouping data into four classes corresponding to a number of athletes between 0 and 9, 10 and 49, 50 and 149, 150 and more. Although the number of participating athletes per nation is not known yet, its evolution across the different Olympics editions does not induce a change of class for any given country with the four above-defined classes; thus a discrete variable for the number of participating athletes (noted RNA below) would fit with forecasting models. Compared to the above explanatory models, the two forecasting models encompass one more explanatory variable: the medal totals four years earlier.

$X_{i,t}\Theta$ is defined by:

$$c + \alpha \ln N_{i,t-4} + \beta \ln\left(\frac{Y}{N}\right)_{t-4} + \gamma Host_{i,t} + \sum_{p} \delta_p PoliticalRegime_{p,i} + \sum_{r} \rho_r \, Regions_{r,i} + \sum_{d} \iota_{nd} \, RNA_{i,d,t} + \theta Maqdisq_{i,t-4} + \lambda Host\ in\ 4\ years_{i,t} + \mu Host\ 4\ years\ ago_{i,t} \quad \mathbf{(3)}$$

The results show that medal totals four years earlier and the different participating athlete classes compared to the class from 0 to 9 athletes have a significant positive impact on medals totals, with an increasing coefficient for the participating athlete classes (Table 3). Compared to the explanatory models, GDP per capita and having hosted the Games four years earlier cease to be significant in the Hurdle model, while fewer political regime and regions dummies are significant in the Tobit model. An explanation is that these four variables are correlated with the medal totals four years earlier, i.e. the variable added in the forecasting models, with it capturing their impact. For GDP per capita, an additional explanation is that it mainly impacts whether a country wins medal(s) or not (rather than the number of medals for countries with at least one medal), i.e. an information provided by the equation not released for the Hurdle model[8]. In the Tobit model, having hosted the Games four years earlier has a significant negative impact. An explanation is that the medal totals four years earlier overestimate the medal total in t for the country having hosted the Games four years earlier. This overestimation is counterbalanced by the dummy variable capturing the fact that the country hosted the Games four years earlier, explaining its significant negative impact.

Table 3

Forecasting National Medals Totals at the 2016 Rio Olympics

Running the two forecasting models based on the results obtained for the 1992-2012 period (not displayed in the paper but available upon request) with the already known data pertaining to the 2016 Games, it appears that they perform well: they are able to predict between 82.3% (Tobit model) and 87.5% (Hurdle model) of overall medal wins with a 95% confidence interval (Table 4). Extending beyond the confidence interval by a two medals error margin, between 91.1% (Tobit model) and 93.2% (Hurdle model) of the distributed medal totals are correctly predicted. The Hurdle model performs better than the Tobit model with a 95% confidence interval. Nevertheless, in a number of cases, its confidence interval is larger and leads to consider a forecast as accurate while this would not be the case with the confidence interval of the Tobit model. This is less frequently the case the other way round, meaning that the Hurdle model is more likely to present a better percentage independently of whether its exact forecasts are better than the Tobit model or not. To try to control for this issue, we calculated what would have been the rate of right forecasts for the Hurdle model with the 95% confidence interval of the Tobit model. Interestingly, the results of the Hurdle model remain better than the Tobit model (87.0% of the distributed medal totals correctly predicted with a 95% confidence interval, 92.2% when the confidence interval is extended by a two medals error margin). Given that the latter is the standard forecasting model since Bernard and Busse (2004) and a Hurdle model has never been tested to forecast national medals totals at Olympic Games, finding that the Hurdle model performs better with a 95% confidence interval is an important contribution to the forecasting literature.

Table 4

With a view to optimise forecasts, it is worth investigating further the differences between the Hurdle and the Tobit models, as well as what works better with one model or the other. If the Hurdle model performs better with a 95% confidence interval, this is not the case for the rate of exact forecasts. Indeed, the Hurdle model forecasts correctly 21.9% of the numbers of medals *vs.* 43.2% for the Tobit model. More exactly, the Tobit model performs better when it comes to forecast which countries end with 0 medal (69.2% *vs.* 21.5% for the Hurdle model), while the Hurdle model performs better when it comes to forecast which countries end with 1 medal and more (22.4% *vs.* 10.6%, including the host country Brazil for the Hurdle model). These elements highlight that both models are complementary.

Forecasting National Medals Totals at the 2020 Tokyo Olympics

Now when forecasting models are run for the 2020 Games they come out with the predictions shown in Table 5. First, it is worth noting that our models forecast exactly the same set of top 13 countries as the one found in Olympic Medals Predictions (2020). Nevertheless, the respective country rankings and number of medals reveal some differences. Both forecasts converge on the United States ending first with a large margin and China ending second. The most striking differences between modelled forecasts and Olympic Medals Predictions show up for Russia (-16 medals in models) and France (+19 to +27).

Medal totals for France heavily depend on whether the variables having hosted the Games four years earlier and hosting the Games four years later (since France is going to host the Games in 2024) are taken on board or not. When they are removed from the models, the forecast for France is 42 or 43 medals, that is a lower medal total than for Japan and Germany. A key factor determining the medal total for France in 2020 would be whether preparing an Olympic team for 2024 had been engaged as soon as in 2017, *i.e.* when Paris was awarded hosting the 2024 Games, and whether such preparation would have a positive impact as early as in 2020.

Table 5

Conclusion

This paper aimed at explaining previous national team medal totals at the 1992-2016 Summer Olympic Games (n = 1289 observations) with a set of variables similar to previous studies, though including the test of a (significant) regional variable which was not taken on board in the literature in English so far. Another objective was to work out econometric testing not only resorting to a Tobit model as usual but also to a Hurdle model. Two explanatory models were then implemented in such a way as to forecast national team medals totals at the 2016 and 2020 Summer Olympics. Forecasting national team medal totals at the 2016 Summer Olympics shows that the Hurdle model performs better than the Tobit model with a 95% confidence interval, questioning the relevance of using (only) the latter that became standard since Bernard and Busse (2004) and, as such, making an important contribution to the literature. Forecasting national team medal totals at the 2020 Summer Olympics provides results that are consistent with Olympic Medals Predictions (2020), although some striking differences are found.

FOOTNOTES

[1] Now, WADA (World Anti-Doping Agency) and its national agencies can ask anti-doping tests during ten years after the Games. Consequently, the final actual outcome of the Games is definitively stabilised only in t + 10 (in 2026 as regard the 2016 Rio Games), and disqualifications may happen at any moment meanwhile.

[2] According to WADA published data, only between 0% and 1.9% of all tested athletes are found positive (doped), depending on which sport discipline they compete in.

[3] Communist countries were outliers in the following sense: for instance the GDR, the USSR, etc., were winning much more Olympic medals than non-communist countries with comparable GDP per capita and population.

[4] Czechoslovakia as regard data for 1992, before the split with Slovakia in 1993.

[5] Republic of Serbia-Montenegro from 1992 to 2006, before the split with Montenegro.

[6] As, for instance, sprint in North America, Jamaica and the Caribbean, marathon and long distance running in Ethiopia, Kenya and Eastern Africa, weightlifting in Bulgaria, Turkey, Azerbaijan, Iran, etc.

[7] The equation related to the probability (Probit model) of not winning one medal and the associated estimations are not reported in this article. The Probit part and the negative binomial model are assumed to be uncorrelated.

[8] Results for the first equation in the Hurdle model are available upon request.

REFERENCES

Andreff, Madeleine, Wladimir Andreff, and Sandrine Poupaux. 2008. “Les Déterminants Economiques de la Performance Sportive : Prévision des Médailles Gagnées aux Jeux de Pékin” [Economic Determinants of Sport Performance: Forecasting Medals Won at Beijing Games]. *Revue d’Economie Politique* 118: 135–69.

Andreff, Wladimir. 2009. “Comparaison entre les Prévisions et les Médailles Gagnées aux Jeux de Pékin” [Comparison between Forecasts and Medals Won at Beijing Games]. Pp. 241–47 in INSEP, ed., *Pékin 2008: Regards Croisés sur la Performance Sportive Olympique et Paralympique* [Beijing 2008: Meetings of Mind on Olympic and Paralympic Sport Performance]. Paris: INSEP, Ministère de la Santé et des Sports, Secrétariat d’Etat aux Sports.

Andreff, Wladimir. 2013. “Economic Development as Major Determinant of Olympic Medal Wins: Predicting Performances of Russian and Chinese Teams at Sochi Games.” *International Journal of Economic Policy in Emerging Economies* 6: 314–40.

Andreff, Wladimir. 2019. *An Economic Roadmap to the Dark Side of Sport*. Cham, Switzerland: Palgrave Pivot in Sports Economics.

Bernard, Andrew B., and Meghan R. Busse. 2004. “Who Wins the Olympic Games: Economic Resources and Medal Totals.” *Review of Economics and Statistics* 86: 413–17.

Blais-Morisset, Paul, Vincent Boucher, and Bernard Fortin. 2017. “L’impact des Dépenses Publiques Consacrées au Sport sur les Médailles Olympiques.” [The Impact of Public Spending Dedicated to Sport on Olympic Medals]. *Revue Economique* 68: 623–42.

Celik, Onur Burak, and Mark Gius. 2014. “Estimating the Determinants of Summer Olympic Game Performance.” *International Journal of Applied Economics* 11: 39–47.

Forrest, David, Ismael Sanz, and J.D. Tena. 2010. “Forecasting National Team Medal Totals at the Summer Olympic Games.” *International Journal of Forecasting* 26: 576–88.

Forrest, David, Ian G. McHale, Ismael Sanz, and J.D. Tena. 2015. “Determinants of National

Medals Totals at the Summer Olympic Games: An Analysis Disaggregated by Sport." Pp. 166–84 in Placido Rodriguez, Stefan Késenne, and Ruud Koning, eds., *The Economics of Competitive Sport*. Cheltenham, UK: Edward Elgar.

Forrest, David, Ian G. McHale, Ismael Sanz, and J.D. Tena. 2017. "An Analysis of Country Medal Shares in Individual Sports in the Olympics." *European Sport Management Quarterly* 17: 117–31.

Hu, Mei-Chen, Martina Pavlicova, and Edward V. Nunes. 2011. "Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial." *The American Journal of Drug and Alcohol Abuse* 37: 367–75.

Leeds, Eva Marikova, and Michael A. Leeds. 2012. "Gold, Silver, and Bronze: Determining National Success in Men's and Women's Summer Olympic Events." *Journal of Economics and Statistics* 232: 279–92.

Lowen, Aaron, Robert O. Deaner, and Erika Schmitt. 2016. "Guys and Gals Going for Gold: The Role of Women's Empowerment in Olympic Successes." *Journal of Sports Economics* 17: 260–85.

Maennig, Wolfgang, and Christian Wellbrock. 2008. "Sozioökonomische Schätzungen Olympischer Medaillengewinne. Analyse-, Prognose- und Benchmarkmöglichkeiten." [Socio-economic Estimations of Winning Olympic Medals: Analysis, Prognosis and Benchmark Possibilities]. *Sportwissenschaft* 38: 131–48.

Noland, Marcus, and Kevin Stahler. 2016. "What Goes into a Medal: Women's Inclusion and Success at the Olympic Games." *Social Science Quarterly* 97: 177–96.

Noland, Marcus, and Kevin Stahler. 2017. "An Old Boys Club no More: Pluralism in Participation and Performance at the Olympic Games." *Journal of Sports Economics* 18: 506–36.

Olympic Medals Predictions (2020). *Olympic Predictions*. Available at http://olympicmedalspredictions.com/.

Otamendi, F. Javier, and Luis Miguel Doncel. 2014a. "Medal Shares in Winter Olympic Games by Sport: Socioeconomic Analysis after Vancouver 2010." *Social Science Quarterly* 95: 598–614.

Otamendi, F. Javier, and Luis Miguel Doncel. 2014b. "By Sport Predictions through Socio-economic Factors and Tradition in Summer Olympic Games: The Case of London 2012." Pp. 125–47 in Panos M. Pardalos, and Victor Zamaraev, eds., *Social Networks and the Economics of Sports*. Berlin: Springer.

Otamendi, F. Javier, and Luis Miguel Doncel. 2018. "Can Economists Beat Sport Experts? Analysis of Medal Predictions for Sochi 2014." *Social Science Quarterly* 99: 1699–732.

Trivedi, Pravin K., and David M. Zimmer. 2014. "Success at the Summer Olympics: How Much do Economic Factors Explain?" *Econometrics* 2: 169–202.

Vagenas, George, and Dimitria Palaiothodorou. 2019. "Climatic Origin is Unrelated to National Olympic Success and Specialization: An Analysis of Six Successive Games (1996-2016) Using 12 Dissimilar Sports Categories." *Sport in Society: Cultures, Commerce, Media, Politics* 22: 1961–74.

Vagenas, George, and Eleni Vlachokyriakou. 2012. "Olympic Medals and Demo-economic Factors: Novel Predictors, the Ex-host Effect, the Exact Role of Team Size, and the "Population-GDP" Model Revisited." *Sport Management Review* 15: 211–17.

TABLES

TABLE 1

Summary descriptive statistics

	Mean	Standard deviation	Minimum	Maximum
Number of medals	4.96	13.53	0	121
Population in millions (t-4)	33.12	124.10	0.01	1350.70
GDP per capita in K$ (t-4)	14.84	17.70	0.07	125.65
Host country	0.01	0.07	0	1
Number of athletes	57.94	100.81	1	646
Political regime				
CAPME	0.83	0.38	0	1
CEEC	0.06	0.24	0	1
POSTCOM	0.11	0.32	0	1
Sub-continent				
North America	0.05	0.23	0	1
North Africa	0.03	0.16	0	1
Sub-Saharan Africa	0.25	0.43	0	1
Asia	0.15	0.35	0	1
Latin and South America	0.15	0.36	0	1
Eastern Europe	0.14	0.35	0	1
Western Europe	0.11	0.31	0	1
Middle East	0.08	0.27	0	1
Oceania	0.05	0.21	0	1

TABLE 2

Estimation results of four explanatory models

	Model (1) - Hurdle			Model (1) - Tobit			Model (2) - Hurdle			Model (2) - Tobit		
	coef		s.d.	coef		s.d.	coef		s.d.	coef		s.d.
Constant	-9.481	***	0.93	-125.377	***	11.89	-6.108	***	0.87	-62.923	***	9.25
Population in log (t-4)	0.558	***	0.04	6.602	***	0.59	0.364	***	0.04	2.678	***	0.43
GDP per capita in log (t-4)	0.243	***	0.06	3.456	***	0.69	0.165	***	0.05	2.156	***	0.55
Host country in 4 years	0.359	***	0.11	9.352	***	2.16						
Host country t	0.519	***	0.11	17.817	***	2.16						
Host country 4 years ago	0.303	***	0.10	11.104	***	2.16						
Number of athletes/10							0.034	***	0.00	1.101	***	0.06
CEEC	1.134	***	0.41	11.419		7.46	0.952	***	0.31	5.644		4.47
POSTCOM	1.020	***	0.33	14.298	**	5.73	0.875	***	0.25	10.349	***	3.49
North Africa	-1.385	***	0.41	-20.733	***	6.80	-0.947	***	0.33	-8.979	**	4.32
Sub-Saharan Africa	-0.886	***	0.33	-18.859	***	4.73	-0.458	*	0.27	-6.472	**	3.05
Asia	-1.510	***	0.30	-22.291	***	5.02	-0.957	***	0.24	-8.860	***	3.09
Latin and South America	-1.169	***	0.32	-16.422	***	4.89	-0.872	***	0.25	-7.899	***	3.04
Eastern Europe	-0.926	**	0.40	-15.448	**	7.18	-0.678	**	0.31	-8.800	**	4.33
Western Europe	-0.047		0.28	-6.117		4.93	-0.024		0.22	-7.165	**	3.03
Middle East	-1.335	***	0.34	-20.464	***	5.10	-0.839	***	0.27	-7.751	**	3.30
Oceania	0.735		0.46	-4.859		6.86	0.390		0.35	-7.591	*	4.52
$g_{i,t}$	-3.374	***	0.24				-3.341	***	0.24			
σ^2_u	0.350	***	0.06	164.685	***	23.67	0.355	***	0.06	51.907	***	8.36
Observations total	554			1289			554			1289		
Observations non censored				554						554		

Notes: *** significant at the 1% level; ** 5% level; * 10% level; c: test statistics associated to the comparison of the models with and without taking into account panel; coef for coefficient and s.d. for standard deviation.

TABLE 3

Estimation results of two forecasting models

	Model (3) - Hurdle			Model (3) - Tobit		
	coef		s.d.	coef		s.d.
Constant	-3.725	***	0.92	-22.507	***	4.57
Population in log (t-4)	0.231	***	0.04	0.700	***	0.19
GDP per capita in log (t-4)	0.067		0.05	0.569	**	0.28
Host country in 4 years	0.336	***	0.10	8.864	***	2.02
Host country t	0.366	***	0.11	12.520	***	2.03
Host country 4 years ago	-0.050		0.11	-4.760	**	2.08
Number of medals (t-4)	0.016	***	0.00	0.897	***	0.02
Athletes [10,50[	0.510	**	0.23	5.126	***	0.67
Athletes [50,150[	0.989	***	0.24	7.394	***	0.87
150 athletes and more	1.559	***	0.27	9.314	***	1.15
CEEC	0.490	*	0.27	0.390		1.40
POSTCOM	0.538	**	0.22	1.828	*	1.07
North Africa	-0.844	***	0.28	-2.101	*	1.29
Sub-Saharan Africa	-0.246		0.25	-1.749	*	1.04
Asia	-0.618	***	0.21	-1.719	*	0.97
Latin and South America	-0.742	***	0.22	-2.705	***	0.96
Eastern Europe	-0.386		0.26	-1.890		1.35
Western Europe	-0.009		0.18	-1.205		0.90
Middle East	-0.520	**	0.24	-1.373		1.07
Oceania	0.288		0.29	-1.772		1.39
$g_{i,t}$	-3.408	***	0.25			
σ^2_u	0.115	***	0.03	26.68	***	1.66
Observations total	529			1232		
Observations non censored				529		

Notes: *** significant at the 1% level; ** 5% level; * 10% level; c: test statistics associated to the comparison of the models with and without taking into account panel; coef for coefficient and s.d. for standard deviation.

TABLE 4

Forecast of Olympic medals for the 2016 Rio Games

		Model (3) - Hurdle			Model (3) - Tobit		
Countries	Number of medals Rio 2016	**Forecast**	Lower CI	Upper CI	**Forecast**	Lower CI	Upper CI
United States	121	**105**	94	115	**99**	95	102
China	70	**106**	95	117	**89**	86	93
Great Britain	67	**48**	43	53	**56**	51	61
Russia	55	**70**	63	77	**69**	66	71
France	42	**38**	34	42	**36**	34	38
Germany	42	**47**	42	51	**44**	42	46
Japan	41	**43**	38	48	**48**	43	53
Australia	29	**37**	31	43	**35**	31	38
Italy	28	**29**	25	33	**29**	28	31
Canada	22	**19**	15	23	**21**	19	24
South Korea	21	**29**	25	32	**31**	29	33
Brazil	19	**19**	15	23	**33**	28	38
Netherlands	19	**20**	17	23	**22**	20	23
Azerbaijan	18	**11**	8	13	**10**	8	13
Kazakhstan	18	**10**	8	13	**11**	9	13
New Zealand	18	**13**	9	17	**14**	10	17
Spain	17	**18**	15	22	**20**	18	22
Rate of right forecasts for 2016							
All countries (192)							
CI to 95% (+ or -2)			88.5% (93.2%)			83.9% (90.6%)	
Exact forecasts (+ or -1)			21.9% (77.1%)			43.2% (74.5%)	
Exact forecasts 0 medal (107 countries)			21.5%			69.2%	
Exact forecasts non 0 medal (85 countries)			22.4%			10.6%	
Countries with at least 3 medals (56)							
CI to 95% (+ or -2)			64.3% (76.8%)			50% (69.6%)	
Exact forecasts (+ or -1)			8.9% (37.5%)			8.9% (30.4%)	

Note: CI = confidence interval.

TABLE 5

Forecast of Olympic medals for the 2020 Tokyo Games

		Model (3) - Hurdle			Model (3) - Tobit		
Countries	Number of medals Rio 2016	**Forecast**	Lower CI	Upper CI	**Forecast**	Lower CI	Upper CI
United States	121	**139**	127	151	**115**	111	119
China	70	**77**	68	86	**70**	67	73
Great Britain	67	**57**	52	62	**64**	62	67
Russia	55	**55**	49	61	**55**	53	57
France	42	**59**	54	64	**51**	46	55
Germany	42	**45**	41	49	**42**	40	44
Japan	41	**47**	42	52	**53**	49	58
Australia	29	**33**	27	38	**29**	26	32
Italy	28	**29**	25	33	**29**	27	31
Canada	22	**20**	17	24	**25**	23	27
South Korea	21	**24**	21	28	**22**	20	24
Brazil	19	**13**	9	16	**15**	10	20
Netherlands	19	**20**	17	23	**20**	19	22
Azerbaijan	18	**12**	10	15	**18**	16	20
Kazakhstan	18	**12**	9	15	**18**	16	20
New Zealand	18	**14**	10	18	**18**	15	21
Spain	17	**18**	15	21	**19**	17	21
Denmark	15	**9**	7	11	**14**	12	16
Hungary	15	**17**	14	20	**16**	14	18
Kenya	13	**12**	10	15	**12**	10	14
Uzbekistan	13	**8**	5	10	**14**	12	15
Cuba	11	**13**	10	17	**14**	11	17
Jamaica	11	**10**	7	12	**10**	8	13
Poland	11	**14**	10	17	**13**	11	15
Sweden	11	**13**	10	15	**13**	11	15
Ukraine	11	**16**	13	20	**14**	12	16

Note: CI = confidence interval.