Impact du manque de public dans la performance des équipes de football avec étude de cas en période de COVID

Ugo ZENNARO & Thomas JUNOY

2023-05-03

Introduction

Souvent on considère qu'en sports l'équipe à domicile est avantagée par la présence de son public qui encourage celle-ci et parfois cherche à déstabiliser les joueurs adverses. Ici nous nous demanderons si la relation causale du public sur les résultats de l'équipe à domicile en football est significative. Pour bien mesurer l'impact de ce soutien il faut conserver les conditions qui définissent l'équipe à domicile et l'équipe à l'extérieure dans notre étude: l'équipe à domicile joue dans son stade et n'a pas de déplacement. Le traitement observé doit n'avoir que pour seul effet: vider les stades. C'est ce que l'on a observé dans la plupart des pays au moment du COVID pendant plus d'un an. Néanmoins on ne peut pas dire que le COVID n'ait pas eu d'autres effets que de vider les stades de football. Ainsi pour éviter les impactes parasites que la COVID a pu avoir sur les résultats en football et se concentrer sur le traitement défini, il sera important d'également observer un championnat témoin qui, même pendant le COVID, n'a pas vidé ses stades.

Pour mener ce travail nous avons donc choisi d'observer 6 championnats de football, dont ceux qui sont appelés les "5 grands championnats européens" qui correspondent aux ligues de football nationales les plus hautes de chacun des pays suivant: Angleterre, Allemagne, Espagne, France et Italie. Ces championnats là seront ici la population qu'ont considérera traitée puisque les politiques nationales ont demandés que les matchs soient joués sans spectateurs. Le sixième championnat est celui de Russie et il constitue la population non traité que l'on observera puisque celui-ci a conservé ses spectateurs même lors du COVID.

Données

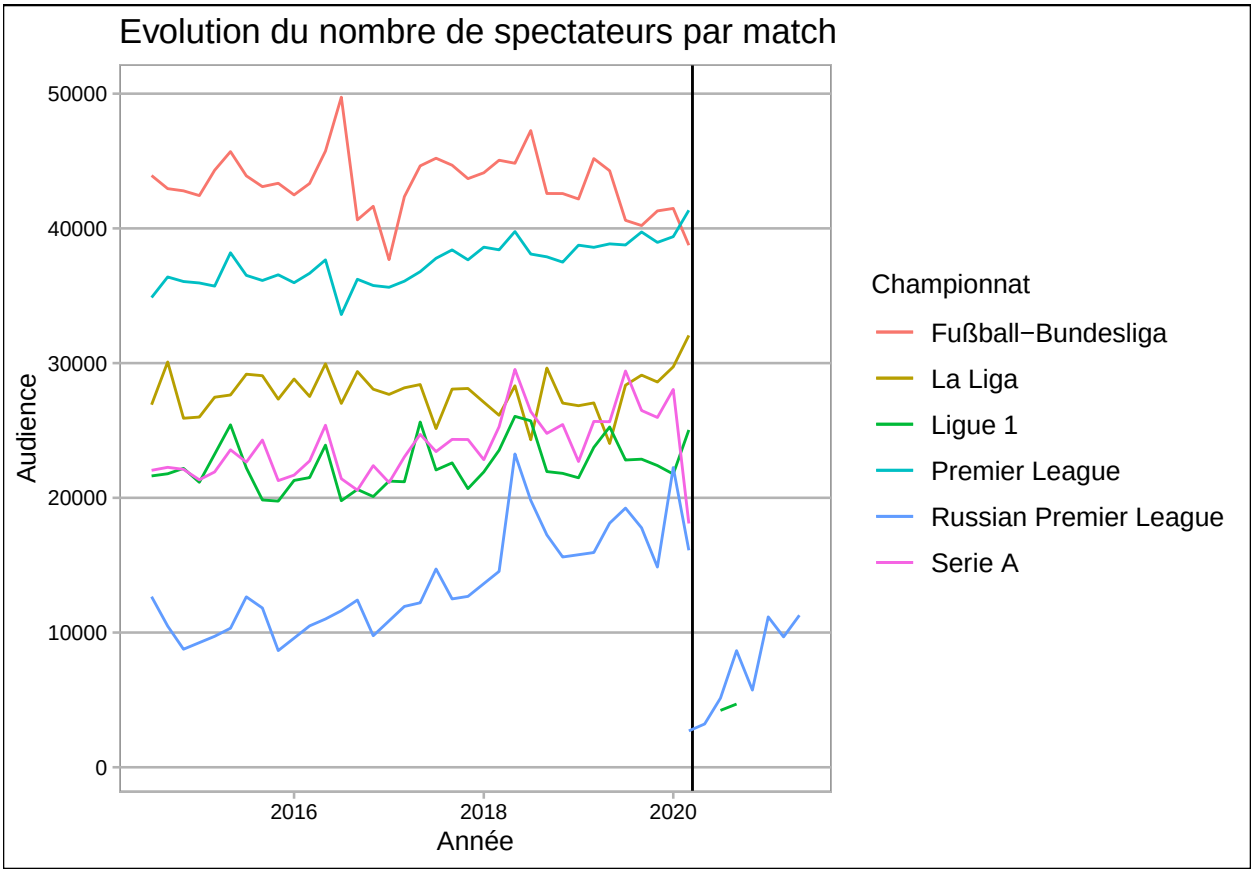
Extraction

Les données utilisées pour ce travail ont été extraites du site `fbref.com` qui regroupe de nombreuses bases de données très riches sur le football. C'est le package `worldfootballR` qui permet d'extraire des données de plusieurs sites très renommés pour leur bases de données relatives au football.

Ici on créée donc deux jeux de données: le premier que concatène les cinq grands championnats de la saison 2014/2015 à la saisons 2020/2021, on aura donc un échantillon sur 7 saisons dont deux durant lesquels l'épidémie de la COVID a eu lieu. Le deuxième sur le championnat Russe sur cette même période. À ces deux jeux de données on marquera la période du COVID à partir du 15/03/2020. Les dates postérieures à cette dates seront considéré post-traitement.

Analyses descriptives

Avant de voir la partie causale de notre étude, on veut s'assurer que le traitement considéré est bien observé en visualisant l'évolution de l'audience moyenne par match au fil des saisons et par championnat.



Cette visualisation préalable est intéressante, on voit dans un premier temps que le championnat Russe est celui dont le nombre de spectateurs moyen est le plus faible que tous les cinq grands championnats mais que l'écart qui les séparait tendait à devenir de plus en plus faible, voire nul, au fil des saisons. Cela n'a pas d'impact sur notre étude mais on peut supposer que cette tendance ait pu avoir un rôle dans la décision de conserver le public dans les stades. Aussi on observe une tentative de retour du public en France pendant le COVID mais pour ce travail nous considérerons cette tentative négligeable tant elle n'a pas durée.

Aussi on peut regarder les résultats moyens de l'équipe à domicile sur le jeu de données observé.

```
## [1] "Victoire :  44.4 % | Nul :  25.2 % | Défaite :  30.4 % | Nombre d'observation :  14391"
```

On observe ici que les matchs sont le plus souvent remportés par l'équipe à domicile et que les défaites sont toutefois plus présentes que les matchs nuls (égalité).

Méthodes

Pour vérifier une relation de causalité entre l'absence de public et la performance de l'équipe à domicile nous observerons plusieurs choses. D'abord nous verrons, par un test de Student si les équipes à domicile marquent plus de buts que l'équipe adverse en regardant si on peut dire que l'hypothèse que l'équipe adverse marque autant ou plus de buts que l'équipe à domicile est rejetable. Aussi en considérant Y_t^0 la population non-traité au temps t et Y_t^1 la population traité au même temps, pour monter l'effet de causalité traitement on doit avoir :

$$\mathbb{E}[Y_t^1] \neq \mathbb{E}[Y_t^0]$$

ce qui revient à :

$$\begin{aligned}\mathbb{E}[Y_t^1 - Y_t^0] &= \mathbb{E}[Y_t^1 - Y_t^0 + (Y_{t-1}^0 - Y_{t-1}^0)] \\ &= \mathbb{E}[(Y_t^1 - Y_{t-1}^0) + (-Y_t^0 + Y_{t-1}^0)] \\ &= \mathbb{E}[Y_t^1 - Y_{t-1}^0] - \mathbb{E}[Y_t^0 - Y_{t-1}^0] \\ &\neq 0\end{aligned}$$

C'est ainsi qu'avec la méthode causale Différences-par-Différences nous chercherons à répondre à notre problématique. Pour dire les choses simplement, l'objectif est ici de comparer l'évolution des observations entre avant et après le traitement chez le groupe traité et chez le groupe contrôle, et en partant du postulat que les deux groupes soient, à l'origine, suffisamment similaires, on quantifiera la relation de causalité par la significativité de l'écart entre les deux évolutions. Ici, il s'agira de voir si, pendant le COVID, l'évolution des performances des équipes à domicile dans les cinq grands championnats change de manière significative par rapport à l'évolution des performances des équipes à domicile en Russie. Pour cela nous utiliserons une régression linéaire qui aura 2 booléens comme variables explicatives de l'écart de performance dû au fait d'être à domicile. Ces deux booléens répondent aux questions suivantes: la population est elle traitée ? la temporalité est elle postérieur au traitement ? La première question permet d'établir un coefficient qui quantifie l'écart entre la population traité et la population contrôle, on verra si cet écart est significatif. La deuxième est associée à l'évolution des performance après traitement mais sur toutes les populations. Aussi nous aurons l'union des deux booléens (le traitement considéré ici) comme variable explicative qui permettra de voir si le coefficient associé est significatif qui permettra d'exprimer si la relation est causale.

Pour tester cela, observons deux indicateurs de performances le taux de victoire de l'équipe à domicile, puis le nombre de buts supplémentaire moyen par match qu'elle marque par rapport à l'équipe adverse (négatif en cas de défaite).

Résultats

Taux de victoire

Tout d'abord, nous pouvons observer par curiosité les taux de victoire dans les 5 grands championnats européens en fonction de la variable d'intérêt `Post-Covid`.

```
## [1] "46% de victoire à domicile avant le covid"
```

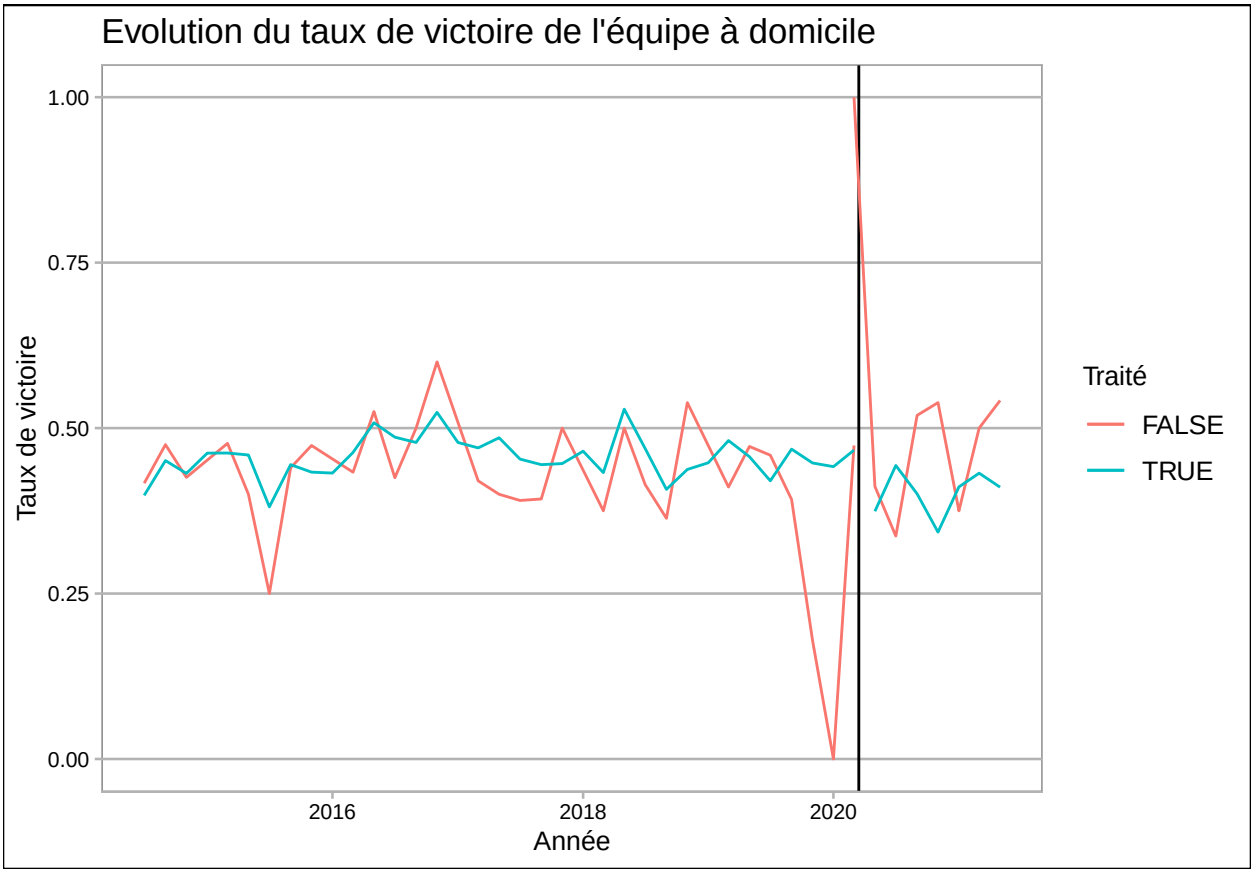
```
## [1] "42% de victoire à domicile durant le covid"
```

À première vue, on observe une baisse de 4% du taux de victoire dans les 5 grands championnats européens. Cela est intéressant à observer, mais il reste à déterminer si cet écart est significatif et s'il existe une causalité entre la présence du public et les résultats des équipes à domicile. Nous allons également examiner ce que cela donne pour le championnat russe.

```
## [1] "42% de victoire à domicile avant le covid"
```

```
## [1] "46% de victoire à domicile durant le covid"
```

La dynamique des équipes à domicile dans le championnat russe semble totalement inversée par rapport aux 5 grands championnats européens. Observons les résultats de la régression linéaire qui exprime la variable booléenne `home_win` selon le fait de jouer en Russie et d'être avant ou après la pandémie de Covid-19.




```
##
## Call:
## lm(formula = home_win ~ (Country != "RUS") * Post_covid, data = C_DATA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4557 -0.4550 -0.4009  0.5450  0.5991
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          0.42444    0.01335  31.793   <2e-16 ***
## Country != "RUS"TRUE                 0.03059    0.01420   2.154   0.0313 *
## Post_covidTRUE                       0.03130    0.03141   0.997   0.3190
## Country != "RUS"TRUE:Post_covidTRUE -0.08544    0.03347  -2.553   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 14387 degrees of freedom
## Multiple R-squared:  0.001691,   Adjusted R-squared:  0.001483
## F-statistic: 8.122 on 3 and 14387 DF,  p-value: 2.116e-05
```

Ces résultats nous permettent de constater plusieurs choses. Tout d'abord, il existe une différence significative (p-value < 0.05) sur le taux de victoire lorsque le pays n'est pas la Russie. Le taux de victoire augmente alors de 3% par rapport à l'estimation de base (42%). L'impact de la variable `Post_covid` ne semble pas significatif selon la régression linéaire (p-value > 0.05). Cependant, lorsque les variables `Country` et `Post_covid` sont croisées (c'est le traitement considéré), l'impact s'avère être significatif avec une p-value sur l'indépendance de 0.01 (<0.05). Le taux de victoire chute alors de 8,5% par rapport à l'estimation initiale du taux de victoire:

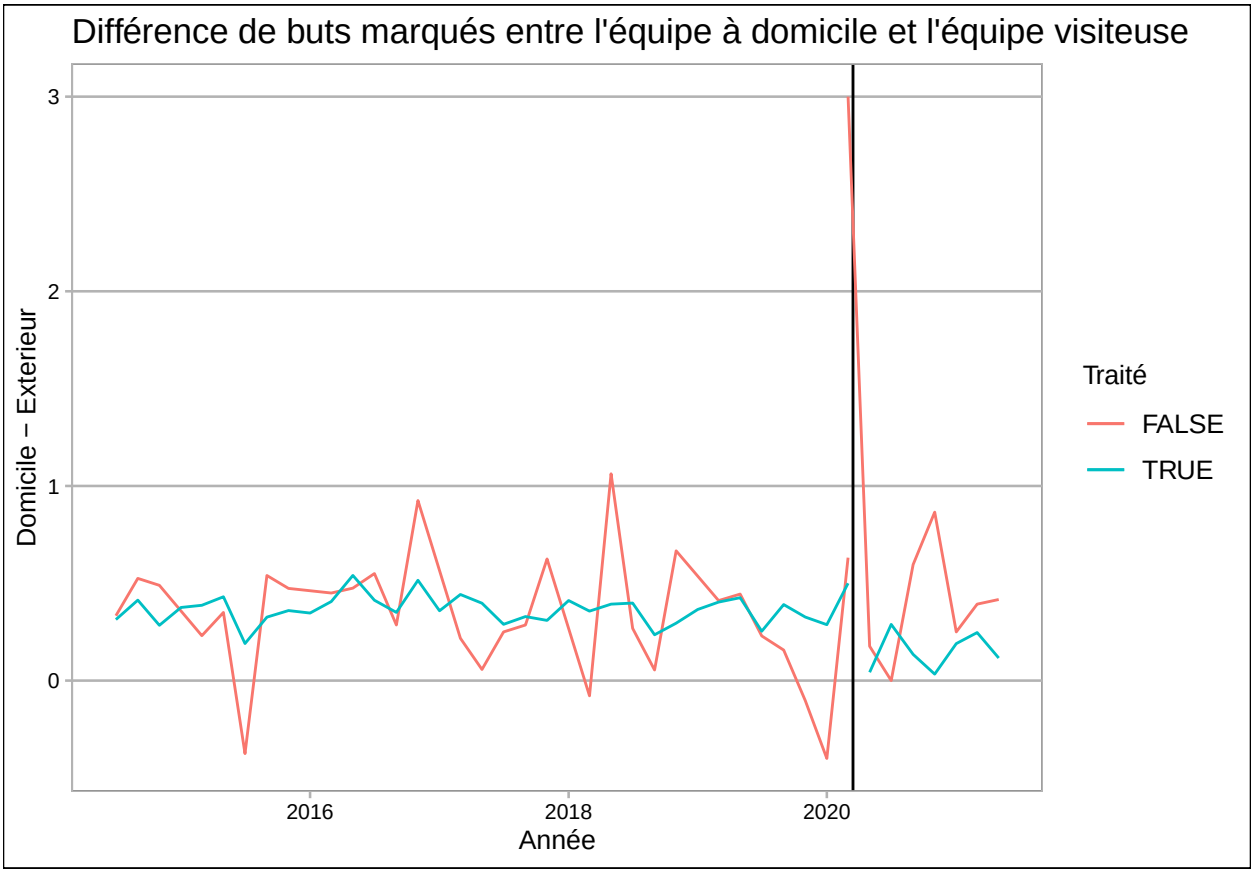
la probabilité estimé qu'aurait l'équipe à domicile de gagner sans publique n'est plus que de 34%. Cette chute semble aussi visible graphiquement.

Différence de buts

Pour commencer, on regarde si les équipes à domiciles marquent plus de buts que les équipes adverses avec un test de Student qui compare le nombre moyen de buts marqués par l'équipe à domicile et le nombre moyen du buts marqués par l'équipe à l'extérieur.

```
##
##  Welch Two Sample t-test
##
## data:  DATASET$HomeGoals[C_DATA$Post_covid == FALSE] and DATASET$AwayGoals[C_DATA$Post_covid == FALSE]
## t = 20.701, df = 20716, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.326273      Inf
## sample estimates:
## mean of x mean of y
##  1.544174  1.189737
```

On lit que le test effectué attribue une probabilité très proche de 0 à l'hypothèse posant que l'équipe visiteuse marque plus ou autant de buts que l'équipe à domicile. Ainsi on rejette cette hypothèse et on peut dire qu'en moyenne les équipe à domicile marquent plus de buts que l'équipe visiteuse. Cela confirme l'intuition que l'on pouvait avoir, mais regardons maintenant si cette avantage fluctue avec le traitement.




```
##
## Call:
```

```
## lm(formula = (HomeGoals - AwayGoals) ~ (Country != "RUS") * Post_covid,
##     data = C_DATA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3620 -1.3620 -0.3138  0.8460  8.8460
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          0.31381    0.04887   6.421  1.4e-10 ***
## Country != "RUS"TRUE                 0.04823    0.05200   0.927   0.3537
## Post_covidTRUE                       0.06652    0.11498   0.579   0.5629
## Country != "RUS"TRUE:Post_covidTRUE -0.27454    0.12251  -2.241   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.818 on 14387 degrees of freedom
## Multiple R-squared:  0.0017, Adjusted R-squared:  0.001492
## F-statistic: 8.168 on 3 and 14387 DF,  p-value: 1.98e-05
```

Premièrement, le graphique et la régression nous confirment que l'équipe à domicile marque en moyenne plus de buts puisque l'`Intercept`, soit l'ordonnée à l'origine de notre régression est un indicateur significatif car la p-value de l'indépendance dans la régression est quasi-nulle (d'ordre 10e-10), de plus il quantifie cet écart avec le coefficient associé qui nous dit qu'en moyenne l'équipe à domicile marque 0.31 buts de plus que l'adversaire. On peut maintenant voir comment cette écart fluctue avec les variables. Si on prends les variables indépendamment l'une de l'autre, on observe des impacts négligeables, être parmi les cinq grand championnats ou pendant le COVID aurait un léger impact positif sur la performance de l'équipe à domicile, mais ici on ne peut pas rejeter l'indépendance de ces variables, on ne peut donc pas dire que celles-ci soient impactantes, cela permet de confirmer la similarité entre les deux populations. Enfin, l'union de ces deux variables, c'est à dire le traitement, semble avoir un impact significatif sur la performance de l'équipe à domicile puisqu'on peut rejeter à 97% l'hypothèse d'indépendance du traitement avec la performance de l'équipe à domicile. Ce traitement semble retirer presque intégralement l'avantage suggéré par l'ordonné à l'origine car et l'avantage ne serait plus que de $0.31 - 0.27 = 0.04$.

Conclusion

En conclusion, nous avons vu que les équipes jouant à domicile sont significativement avantagées. Les spectateurs sont en partie la cause de cet avantage puisqu'être à domicile sans spectateur est beaucoup moins avantageux. Nous avons pu voir cela en observant le taux de victoire à domicile mais également le nombre moyen de buts supplémentaire que l'équipe à domicile a marqué par rapport à son adversaire. On a donc pu observer la réponse à notre question causale: oui la présence du publique a un effet significatif sur les performance de l'équipe à domicile. C'est en ayant isolé cette caractéristique qui découle de décisions politiques dans certains pays liées au COVID, et au fait que la Russie n'ait pas pris la même décision, que nous avons pu répondre à cette question. L'épidémie de COVID a été mondiale, c'est pour cela que nous avons jugé que l'impact sur la Russie était égale, hors traitement observé, aux autres pays, on ne l'a néanmoins ici pas démontré. Son impact étant multifactorielle et ayant occasionné beaucoup d'effet rebonds qui ont pu différer selon les pays, démontrer le postulat que nous avions posé aurait relevé d'une quantité de données et d'un travail bien plus conséquents.