Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 17, No. 1, 2025

Data Structure Trend Prediction Based on XGBoost and DFR Integration

Zhiyi Peng †, Xiangyu Gu †, *

College of Electronic and Optical Engineering & College of Flexible Electronics (Future Technology), Nanjing University of Posts and Telecommunications, Nanjing, China

* **Corresponding author Email:** b23090526@njupt.edu.cn

† These authors also contributed equally to this work

Abstract: This study proposes a predictive modeling framework based on XGBoost and ensemble learning techniques to forecast national outcomes and structural trends in large-scale competition datasets. First, a multi-dimensional feature system incorporating static, dynamic, and interaction variables is constructed to characterize country-specific attributes. Based on this, an XGBoost-based nonlinear regression model is developed to estimate both gold and total medal counts, while integrating host country effects and event-scale adjustments. Next, to address prediction challenges for countries with sparse historical performance, a hierarchical clustering strategy is implemented to identify potential first-time medal winners and assess rank shifts. Finally, a hybrid Difference-in-Differences and Random Forest Regression (DFR) model is introduced to evaluate the quantitative effect of external interventions, with a focus on marginal contribution estimation and residual analysis. Through these methods, the framework achieves improved predictive accuracy, enhanced interpretability of event-level influence, and a refined understanding of competition structure dynamics.

Keywords: XGBoost; DFR Model; Event Importance.

1. Introduction

Accurate forecasting of outcomes in large-scale competitive scenarios necessitates robust analytical frameworks capable of processing complex, high-dimensional datasets. Traditional predictive models often suffer from limitations such as overfitting, low tolerance to nonlinear dependencies, and difficulties in modeling sparse or incomplete data[1,2]. These challenges are particularly pronounced when predicting outcomes for entities lacking rich historical information, resulting in diminished predictive reliability and interpretability. To address these shortcomings, this study proposes a data-driven modeling approach that integrates structural feature engineering with advanced machine learning algorithms, aiming to enhance both forecast precision and model robustness[3].

The framework begins with the construction of multi-dimensional national-level feature vectors, integrating static attributes, recent dynamic trends, and cross-feature interactions. Based on these inputs, a nonlinear regression model is built using the XGBoost algorithm, which has demonstrated strong performance in structured data scenarios and is known for its interpretability and speed[4]. Subsequently, a hierarchical clustering strategy is employed to enhance outcome prediction for data-sparse entities by leveraging inter-entity similarity patterns [5]. Finally, a Difference-in-Differences and Random Forest Regression (DFR) hybrid model is developed to quantify the impact of external interventions and assess event-level contribution variances. This comprehensive framework not only improves predictive accuracy, but also supports scenario-sensitive modeling and enhances generalization capability across diverse application domains[6].

2. Medal count prediction and structural analysis based on XGBoost

To establish the mapping relationship between national characteristics and medal performance, a multi-dimensional feature system comprising static, dynamic, and interaction variables is constructed. Based on these features, country-level input vectors are generated. The XGBoost model is then employed to predict the number of gold medals and total medals, with host country indicators and event scale variables incorporated for structural correction.

Upon completion of model training, the framework is applied to the prediction of medal outcomes for the 2028 Olympics. Ranking shift indices and hierarchical clustering methods are utilized to estimate the likelihood of first-time medal acquisition and to analyze structural changes in global performance patterns. Finally, the model outputs are visualized and a scoring mechanism is introduced to reveal the distribution of dominant events across nations. The modeling process and analytical results are detailed in the following sections of this chapter.

2.1. Data description and feature processing

To support the construction of the medal prediction model, relevant national-level variables are systematically organized. Based on their intrinsic properties, the features are categorized into three types: static features, dynamic features, and interaction features. Each category is represented as a feature vector, which are ultimately merged into a unified national input vector for use in the predictive model.

1). Static features

Static features refer to variables that exhibit limited variation over the prediction horizon. The static feature vector is denoted as:

$$X_{\text{static}} = [g, p, m_h, a_h, s_h, o_h] \quad (1)$$

Here, g and p represent the logarithmic values of a country's GDP and population size, respectively. The variable m_h denotes the total number of historical medals, composed of a_h (athletics), s_h (aquatics), and o_h (other sports).

2). Dynamic features

Dynamic features capture recent performance trends during the most recent Olympic cycles. The dynamic feature vector is defined as:

$$X_{\text{dynamic}} = [m_{t-1}, m_{t-2}, r_m, e_t, h_t, h_g] \quad (2)$$

In this vector, m_{t-1} and m_{t-2} represent the medal counts in the previous and penultimate Olympic Games, respectively. The medal growth rate r_m is calculated as:

$$r_m = \frac{m_{t-1} - m_{t-2}}{m_{t-2}} \quad (3)$$

The variable e_t denotes the total number of events in the current Olympics, h_t indicates whether a country is the host, and h_g reflects the influence of hosting on event variation.

3). Interaction features

To capture the interdependence among features, an interaction feature vector X_{inter} is constructed as follows:

$$X_{\text{inter}} = [g \cdot m_h, e_t \cdot p, h_t \cdot m_h, p \cdot g] \quad (4)$$

Finally, all three feature vectors are concatenated to form the national core feature vector X_{ti}:

$$X_t^i = [X_{\text{static}}, X_{\text{dynamic}}, X_{\text{inter}}] \quad (5)$$

This vector serves as the primary input to the predictive modeling framework.

The feature construction process is illustrated in Figure 1.

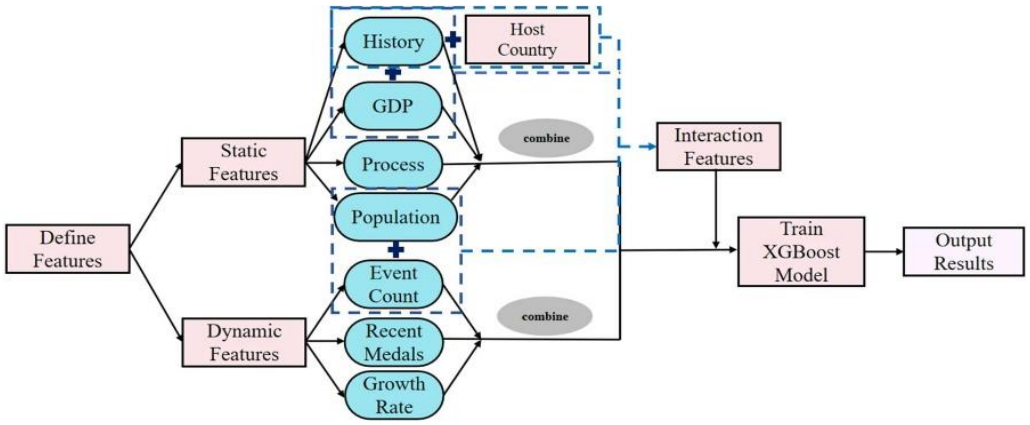

Figure 1. Process of predicting medal list in 2028 Olympics

2.2. Construction of a nonlinear regression model for medal prediction based on XGBoost

1). Basic structure of gold medal and total medal prediction

Based on the constructed input feature vector, two nonlinear predictive functions are formulated to estimate the number of gold medals (y_{ti}) and total medals (z_{ti}), respectively. The expressions are defined as follows:

$$\begin{aligned} y_t^i &= f\left(X_t^i, \theta\right) + \alpha \cdot h_t + \varepsilon_t^i \\ z_t^i &= g\left(X_t^i, \theta\right) + \beta \cdot h_t + \mu_t^i \end{aligned} \quad (6)$$

Here, $f(\cdot)$ and $g(\cdot)$ represent the nonlinear functions trained using the XGBoost algorithm, and X_{ti} is the national feature input vector. The term θ denotes the parameter vector of the model. The binary variable h_t indicates whether the country is the host of the Olympics. The coefficients α and β represent the marginal effect of the host status on the gold medal and total medal counts, respectively. The terms ϵ_{ti} and μ_{ti} refer to the model residuals.

2). Marginal contribution modeling of event scale

Beyond national-level factors, the overall number of Olympic events exerts a significant impact on the distribution of medals. To quantify this influence, a marginal contribution coefficient γ is introduced as the weight of e_t, representing the event count. Specifically, γ_1 and γ_2 denote the marginal contributions of the number of events to the gold medal count and total medal count, respectively. The computation of γ is given by:

$$\gamma = \frac{\text{Cov}(z, e_t)}{\text{Var}(e_t)} \quad (7)$$

The corresponding definitions for covariance and variance are as follows:

$$\text{Cov}(z, e_t) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(e_i - \bar{e})^2 \quad (8)$$

$$\text{Var}(e_t) = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2 \quad (9)$$

In Equation (7), the denominator reflects the variance of e_t, indicating its dispersion, while the numerator represents the covariance between total medal count z and the number of events e_t.

To account for the additional events introduced by the host nation, a variable a_t is defined to represent the number of newly added events. Adjustment terms ΔY_t^{host} and ΔZ_t^{host} are introduced to capture the impact of these new events on

the gold medal and total medal counts, respectively. The expressions are as follows:

$$\Delta Y_t^{host} = \varphi_1 \cdot a_t + \delta_1$$
$$\Delta Z_t^{host} = \varphi_2 \cdot a_t + \delta_2 \quad (10)$$

Here, ϕ_1 and ϕ_2 denote the marginal effect of newly added events on the respective response variables, while δ_1 and δ_2 are residual terms subject to minimization during model training.

Based on the inclusion of the above adjustment variables, the original prediction functions (6) are reformulated as:

$$y_t^i = f\left(X_t^i, \theta\right) + \alpha \cdot h_t + \gamma_1 \cdot e_t + \Delta Y_t^{host} + \varepsilon_t^i$$
$$z_t^i = g\left(X_t^i, \theta\right) + \beta \cdot h_t + \gamma_2 \cdot e_t + \Delta Z_t^{host} + \mu_t^i \quad (11)$$

These expressions represent the final predictive equations for the gold medal count and total medal count. Once the model is constructed and prediction results are obtained, the complete medal table can be generated accordingly.

During model training, hyperparameters such as learning rate and tree depth are iteratively optimized. The final configuration sets the learning rate at 0.1 and the maximum tree depth at 3. Each prediction result is accompanied by a 95% confidence interval, computed as:

$$\hat{y}_t^i \in \left[y_t^i - 1.96\sigma_{\varepsilon_t^i}, y_t^i + 1.96\sigma_{\varepsilon_t^i} \right]$$
$$\hat{z}_t^i \in \left[z_t^i - 1.96\sigma_{\mu_t^i}^i, z_t^i + 1.96\sigma_{\mu_t^i}^i \right] \quad (12)$$

Here, $\sigma_{\epsilon_{ti}}$ and $\sigma_{\mu_{ti}}$ denote the standard deviation of residuals for gold medal and total medal predictions, respectively, capturing the range of model uncertainty.

3). Construction of ranking shift Indicator and first-time medal probability estimation

To capture the projected ranking dynamics of countries in future Olympic Games, a ranking shift indicator ΔR_{34}^i is defined to measure the change in rank between the 34th Olympic Games (Los Angeles) and the 33rd Olympic Games (Paris). The indicator is formulated as:

$$\Delta R_{34}^i = R_{34}^i - R_{33}^i \quad (13)$$

Here, R_{34}^i represents the predicted rank of country i in the 34th Olympics, while R_{33}^i denotes its actual rank in the 33rd Olympics. By comparing ΔR_{34}^i to zero, the direction of ranking change for each country can be determined. The interpretation is summarized in table 1.

Table.1. Different ΔR_{34}^i values and meanings

Symbol	Meanings
$\Delta R_{34}^i > 0$	Rank up
$\Delta R_{34}^i = 0$	Rank remains
$\Delta R_{34}^i < 0$	Rank down

A hierarchical clustering method is applied to refine XGBoost predictions by mapping countries with missing data to structurally similar profiles, improving estimates of first-time medal probabilities.

2.3. Medal prediction results and structural analysis of dominant events

After computation, refinement, and standardized representation, prediction results were obtained for each country, enabling consistent cross-national comparison. A rank-shift indicator was constructed to evaluate performance trends, revealing that 22 countries are projected to improve their rankings, while 20 are expected to decline. Representative changes in national rankings are shown in table 2.

Table.2. Rank changes of representative nations or areas

Nation/area	Changes	Nation/area	Changes	Nation/area	Changes
Italy	$3\uparrow$	China	0	UK	$-1\downarrow$
France	$1\uparrow$	Germany	0	Zambia	$-1\downarrow$

As shown in Figure 2, the top-ranked countries are projected with corresponding gold and total medal counts, accompanied by 95% confidence intervals represented by black lines. The average prediction errors for gold and total medals are 0.5 and 1.7, respectively, indicating high model accuracy.

To quantify the dependence of each country on specific event types, a heatmap was constructed to visualize event-level importance weights. As shown in Figure 3, darker regions indicate higher relative importance.

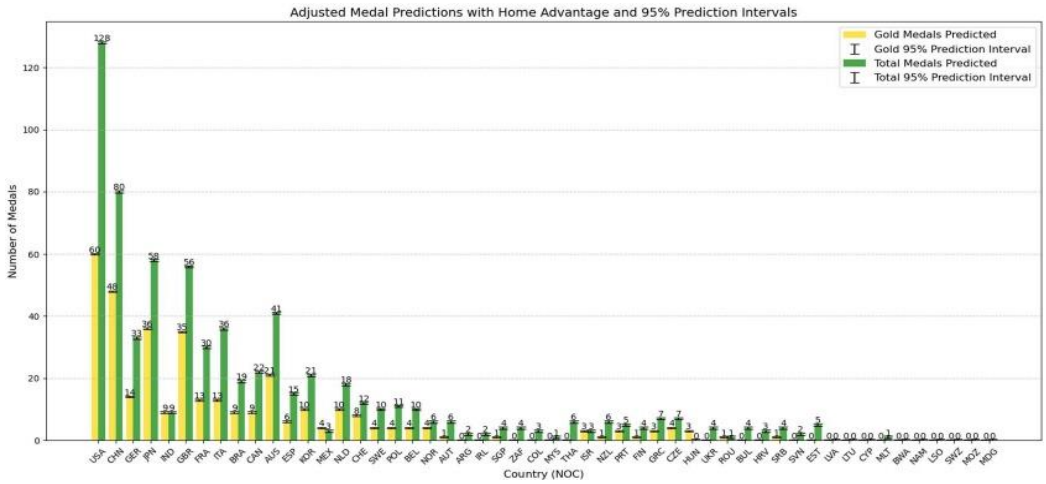

Figure 2. Prediction of numbers of gold medals and medals of each country obtained in LA Olympics with interval

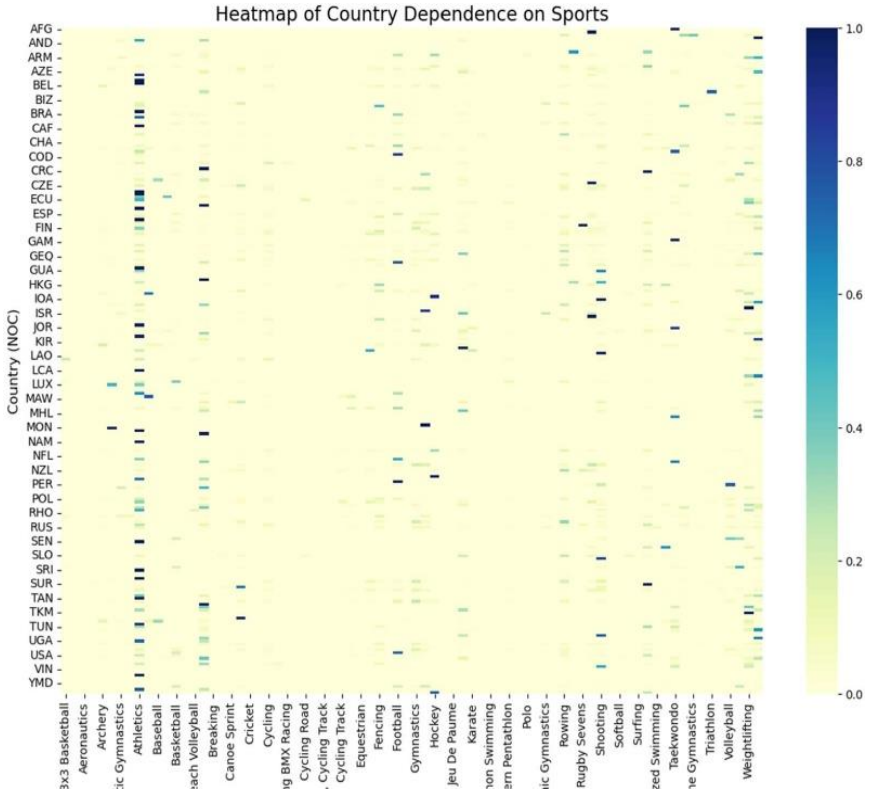

Figure 3. Neural network structure

3. External intervention modeling and strategy optimization via DID-RFR

3.1. Difference-integrated modeling of intervention effects (DFR Model)

This section introduces the DFR model (Difference-in-Differences & Random Forest Regression), which integrates the Difference-in-Differences (DID) method and Random Forest Regression (RFR) to quantitatively evaluate the effect of introducing high-impact external interventions at the national level. The DID component enables robust causal inference on a macro scale, suitable for measuring the overall impact of the intervention on medal outcomes. In contrast, the RFR model captures nonlinear responses at the event level and quantifies the marginal contributions of individual variables.

To evaluate the effect of the intervention, the variables and their corresponding definitions are listed in Table 3.

Table 3. Different variables and meanings to curve the effect

Symbol	Meanings
i	Number of a certain country
j	Number of a certain and independent sport event
Z_i^t	Total medals in No.t Olympics of a certain country
$Z_t^{i,j}$	Total medals in an event in No.t Olympics of a certain country
C_t^i	Index of introduce of great coach
n_j	Number of sub-events in a certain event
P_j^i	Potential for a certain country in a certain event
S_j	Medal density for a certain event

In this context, $C_t^i = 1$ indicates that country i introduced an elite coach during the t-th Olympic Games; otherwise, $C_t^i = 0$. The variable P_j^i represents the average number of medals won by country i in event j over the past five Olympic Games, serving as an indicator of performance potential. For model simplification, only major representative events are considered. The medal density S_j quantifies the proportion of medals attributed to event j and is calculated as follows:

$$S_j = \frac{\sum_{j=1}^{n_j} \text{number of medals in a sub} - \text{event}}{n_j} \tag{14}$$

Following the logic of the DID model, a treatment group ($C_t^i = 1$) and a control group ($C_t^i = 0$) are defined, and the change in medal count is estimated using the following formulation:

$$\Delta Z_t^i = \omega + h \cdot C_t^i + q \cdot g \cdot + r \cdot p + \phi_t^i \tag{15}$$

Here, g and p denote the logarithmic values of national economic scale and population, h and q are their corresponding weights, ω represents the baseline medal count under controlled conditions, and ϕ_t^i is the residual error term.

At the event level, the RFR model integrates multiple decision trees to enhance prediction accuracy and applies feature importance analysis to quantify the effect of coach introduction. The model's predictive vector is defined as:

$$Z_t^{i,j} = RF\left[C_t^i, f\left(X_t^i, \theta\right), S_j, C_t^i, P_j^i, C_t^i \cdot S_j, C_t^i \cdot P_j^i\right] \tag{16}$$

To reflect the effect of coaching on event outcomes, two interaction terms are introduced: $C_t^i \cdot S_j$ for the joint effect

of coaching and project density, and $C_t^i \cdot P_j^i$ for the interaction between coaching and national potential.

3.2. Predictive analysis of effective intervention outcomes

Success probability increases and residual reductions confirm that coaching intervention significantly improves performance consistency. As shown in Figure 4, the average probability of success across events rises by approximately 40% with coaching, indicating its substantial effect on outcome enhancement.

Figure 5 further compares the distribution of residuals under both conditions—before and after the inclusion of a coach. The results indicate a significant reduction in residual values, from 100 to 10, demonstrating enhanced prediction stability. This evidences that the integration of coaching not only improves athletes' actual performance but also contributes to the overall robustness of the forecasting model.

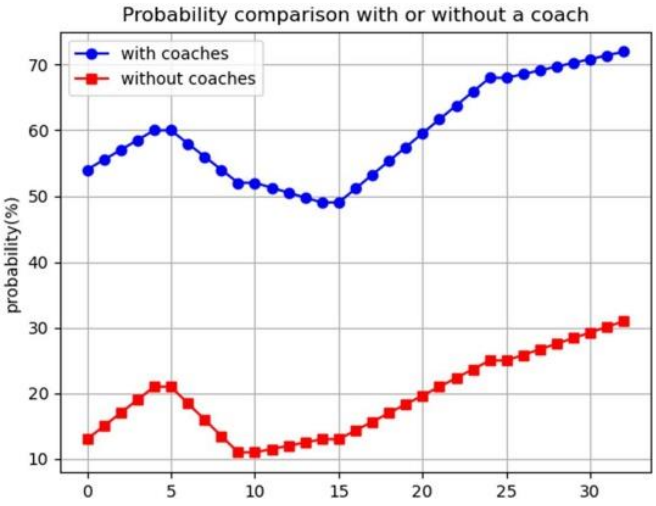

Figure 4. Possibility shifts between coaches introduced and without coaches

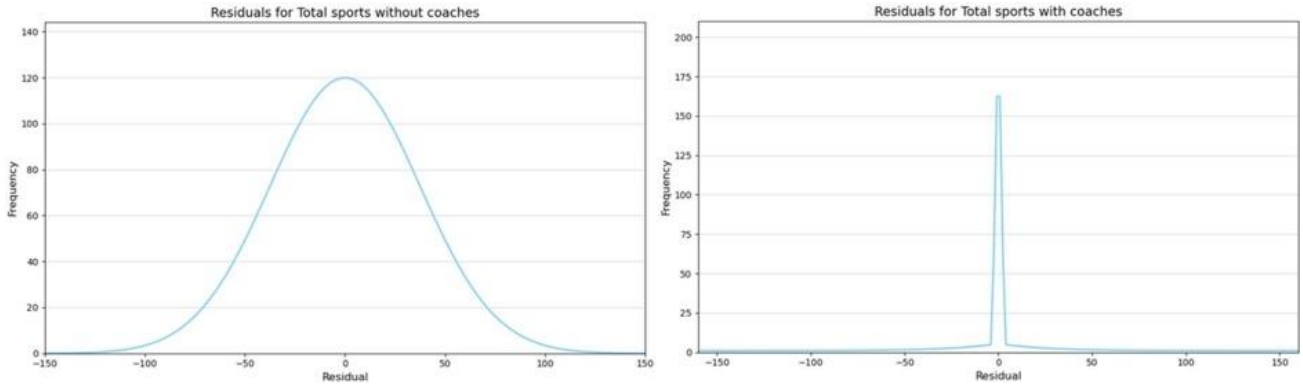

Figure 5. Residual changes between coaches introduced and without coaches.

4. Conclusions

This study establishes an integrated modeling framework based on XGBoost and DFR algorithms to investigate the prediction and structural estimation of large-scale multi-dimensional data systems. First, a nonlinear regression model driven by XGBoost was constructed to address the problem of outcome prediction under complex feature interactions. Through iterative optimization of model parameters and vectorized representation of static, dynamic, and interaction attributes, the framework achieved high accuracy and generalization ability across varying input conditions. Second, to cope with the uncertainty inherent in sparse data environments, a hierarchical clustering strategy was employed to enhance the prediction stability for data-deficient instances and to identify entities with potential structural changes. Finally, a hybrid Difference-in-Differences and Random Forest Regression model was designed to quantify the marginal influence of external interventions. Feature interaction terms were introduced to improve the model's sensitivity and explanatory power. Overall, the proposed approach demonstrates reliable predictive performance, interpretable variable importance, and adaptability to high-dimensional forecasting tasks. Future work may explore integration with deep learning architectures to further enhance scalability and precision.

References

[1] Zhang, H., Wang, Y., & Qian, H. (2021). Limitations of linear models in high-dimensional prediction tasks. Applied Intelligence, 51(6), 3795–3809.

[2] Bekkerman, R., Bilenko, M., & Langford, J. (2011). Scaling up machine learning: Parallel and distributed approaches. Cambridge University Press.

[3] Dietterich, T. G. (2000). Ensemble methods in machine learning. In International Workshop on Multiple Classifier Systems (pp. 1–15). Springer.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD (pp. 785–794).

[5] Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. Journal of the American Statistical Association, 105(490), 713–726.

[6] Abadie, A. (2005). Semiparametric difference-in-differences estimators. The Review of Economic Studies, 72(1), 1–19.