Adoption of Genetically Engineered Seeds in China: Predicting Treatment Effects on the Crop-Yield Distribution by Synthetic Control

Yuansen Li
Department of Agricultural Economics
Purdue University
Email: li4238@purdue.edu

Tor Tolhurst
Department of Agricultural Economics
Purdue University
Email: ttolhurs@purdue.edu

Selected Paper prepared for presentation at the 2024 Agricultural & Applied Economics Association Annual Meeting, New Orleans, LA; July 28-30, 2024

Introduction

In late 2023, China approved its first genetically engineered (GE) corn and soybean seeds. China began breeding GE crop seeds in 1986 and licensed Bt-cotton for commercial use in 1997. However, subsequent controversies slowed development, approval, and adoption. By 2019, the only widely adopted seed was GE cotton. Imports of agricultural products produced with GE technologies in foreign countries were effectively banned. For example, imports of GE alfalfa were approved only in January 2023, ten years after applications were submitted. In 2020, for the first time, the Central Economic Conference emphasized the importance of GE seeds. Indeed, Chinese officials have been supportive of the technology, particularly as a means of strengthening food security. Yet, approvals for human consumption remain forthcoming.

The policy debate on GE is complicated by its uncertain effect on crop yields. In particular, GE's effect on the higher moments of yields remains an open question. In this project, we measure the treatment effect of GE adoption on crop yield distribution in China. We focus on corn, the world's second-largest crop by value FAO (2016). The ideal experiment would measure changes in the distribution of crop yields over time across a treatment regime that has access to GE and a control regime that does not. To mimic this design, we use the recently introduced distributional synthetic control (DSC) estimator Gunsilius (2023). We use province-level data from the US, Argentina, and Brazil (GE regimes) to create a synthetic counterfactual for provinces in China (non-GE regime). We can then estimate the distributional effects of GE on Chinese yields as the difference between the observed yield distribution and the synthetic control estimate. Analogous to a conventional synthetic control, the key identifying assumption is that the weights used to estimate the synthetic distribution are constant across the pre- and post-period.

In this paper, we will first introduce the empirical models we used. Meanwhile, the empirical process will be discussed. Furthermore, we will examine our estimates.

literature review

Lusk, Tack, and Hendricks (2018) utilizes variations in the adoption rates of GE corn across U.S. counties to estimate the heterogeneous treatment effect of adopting GE corn using a panel data method, where the authors control for weather and regional fixed effect. This paper only considers the treatment effect on the conditional mean yield rather than other moments. Scheitrum, Schaefer, and Nes (2020) estimates the long-run GE effects across different countries, but the paper doesn't assume the heterogeneous treatment effects spatially. The econometric model in our paper will properly handle the heterogeneous treatment effects using the Synthetic Control Method (SCM).

SCM has been the most crucial innovation in the policy evaluation literature in the last 15 years Athey and Imbens (2017).Abadie, Diamond, and Hainmueller (2007) estimated the treatment effect of tobacco tax on cigarette consumption in Califonia using SCM. Following this, the SCM has been widely used, yet most empirical studies didn't consider the treatment effect on higher moments. Technically, the classical SCM is to find the shortest distance between a convex hull and the point outside this convex combination. What if all the vertices of the convex polytope and the point outside this convex combination are random variables? Therefore, the fundamental question of DSC lies in finding a way to measure the distance between two random variables or two distributions.

Chen (2020) proposed a method to measure the distance between two random variables and then mimic the behavior of a target distribution using the mixture of multiple distributions. However, this paper gave weights to specific quantiles rather than the whole control unit. Therefore, the method is sensitive to the choice of bins. The Bayesian Model Averaging (BMA) method in Ker, Tolhurst, and Liu (2016) could be a potential candidate to extend the classical SCM into the distributional space. However, the BMA method performs poorly when there is a low degree of overlap in the support of the probability density functions. DSC estimator in Gunsilius (2023) uses the 2-Wasserstein distance to address these problems. This metric space is flexible, so it can provide a reasonable measurement even if the

supports of two distributions are poorly overlapped. Meanwhile, this method can find the optimal weight for the entire quantile function, mitigating the influence of the choice of bins.

The other focus of this paper pertains to estimating the year-conditional yield density of corn. Tolhurst and Ker (2015) employs a mixture of two normal distributions to establish a statistical model for yield density. Meanwhile, it utilizes the Expectation-Maximization (EM) algorithm to estimate the time-embedding trend and the heteroskedastic variance within the distribution. This article utilizes a mixture of two Gaussian distributions to simulate yield density. A limitation is identified in the arbitrary selection of the number of Gaussian distributions within the mixture (to reduce the computational complexity), reducing the model's flexibility. To enhance the flexibility of the Two-Gaussian mixture model and to account for the similarities in climate and soil quality across adjacent corn-producing regions, Ker, Tolhurst, and Liu (2016) employs BMA to estimate corn yield density at the county level in the United States. In other words, the article leverages a weighted average of multiple two Gaussian mixture models to accommodate time-conditional yield density.

Building upon the foundation of Ker, Tolhurst, and Liu (2016), Schuurman and Ker (2024) incorporates climatic data into the BMA framework. This article employs a neural network with one hidden layer to learn the linkage between climatic variables and the weights of the sub-models within the Gaussian mixture. Additionally, it introduces a penalizing term like Ridge Regression and utilizes cross-validation to select the optimal number of sub-models in the mixture. The paper continues to use Gaussian distributions as an underlying assumption, and whether the model outperforms non-parametric methods remains an open question. In our paper, we do not directly estimate time-conditional yield density. Instead, we employ quantile regression to identify quantile functions conditional on time. This approach enables us to obtain empirical quantiles for each year and region. In other words, we acquire information about the distribution without making Gaussian assumptions. The feasibility of applying this indirect methodology lies in the DSC algorithm Gunsilius

(2023); empirical quantiles can be directly incorporated in the DSC, thus preventing the requirement for direct estimation of yield density.

Empirical Model

Quantile regression

We can only obtain time series data for corn yield for each state or province in China, Argentina, Brazil, and the United States. Only one data point can be observed for a specific year rather than the potential yield distribution. In order to obtain the distribution information of potential corn yields for a specific region in a specific year and to better fit with the DSC method, we use quantile regression to obtain the empirical quantiles for a region in a given year.

$$y_t = \alpha_0 + \alpha_1 t + \epsilon_t \tag{1}$$

$$y_t = \beta_0 + \beta_1 t + \beta_3(\max(0, t - t_{T_{0-1}})) + u_t \tag{2}$$

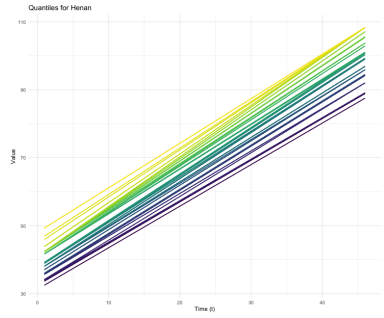
Equations (1) and (2) show the specifications of the quantile regression for the units in the treated group (China) and control groups (Argentina, Brazil, and the United States), respectively. y_t denotes the corn yield for the t-th year. In equation (2), T_{0-1} represents the year before the first introduction of GE seeds of corn. For instance, The United States officially approved the use of genetically modified corn seeds in 1996, so T_{0-1} for the US is 1995. There are two considerations for using equation (2) on the units in the control groups (GE-regimes). Firstly, we don't want the information after treatment to influence the estimation of quantile curves before treatment because the yield after treatment will contaminate the weights selection before treatment, which is the classical setting in the SCM Abadie, Diamond, and Hainmueller (2007). Secondly, we want to incorporate the information before treatment into the estimation of the quantile curve after treatment and

estimate the conditional quantile curves of the GE-regimes simultaneously.

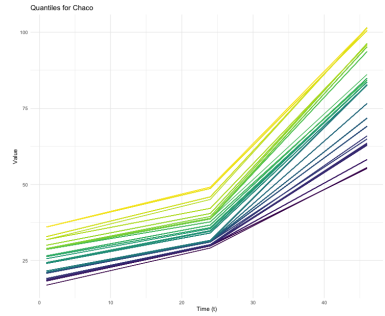
$$\min_{\alpha} \sum \rho_\tau(y_t - \alpha_0 + \alpha_1 t) \tag{3}$$

$$\min_{\beta} \sum \rho_\tau(y_t - [\beta_0 + \beta_1 t + \beta_3(\max 0, t - t_{T_{0-1}})]) \tag{4}$$

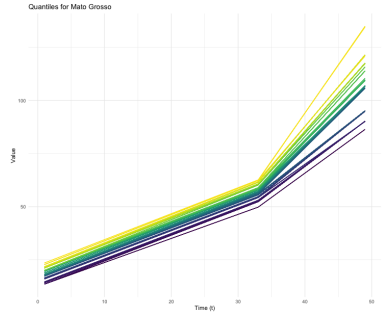
Equations (3) and (4) show the optimization problems related to the specifications defined in equations (1) and (2), respectively, where the ρ_τ denotes the check function for the quantile regression at the τ-th quantile Koenker and Hallock (2001). Meanwhile, in order to get the valid empirical quantile curves (non-crossing), we utilize the method outlined in Muggeo et al. (2013) to obtain the non-crossing quantile. Figure 1 shows the quantile curves of four regions from four countries. In the DSC estimators, we use the empirical quantiles obtained from the conditional quantile curves as the input data.



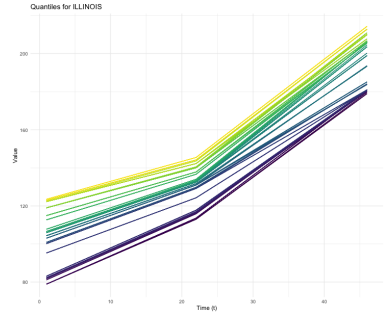
(a) Quantile curves of Henan in China.



(b) Quantile curves of Chaco in Argentina.



(c) Quantile curves of Mato Grosso in Brazil.



(d) Quantile curves of Illinois in the U.S.

Figure 1: Corn yield quantile curves from four regions.

Distributional Synthetic Control

After getting the empirical quantiles of the conditional corn-yield distribution both in GE regimes and non-GE regimes, we will utilize the DSC method to estimate the GE treatment effect on conditional yield density. The treatment units are all the provinces in China; the control units are all the states or provinces in Argentina, Brazil, and the US. The following method to construct the counterfactual distribution is from Gunsilius (2023).

$$W_2(P_1, P_2) = \left(\int_0^1 \left| F_1^{-1}(q) - F_2^{-1}(q) \right|^2 dq \right)^{1/2} \tag{5}$$

Equation (5) defines the 2-Wasserstein distance between the distribution P_1 and P_2. $F_i^{-1}(q)$, where $i \in \{1, 2\}$, denotes the quantile function for the distribution i. Intuitively, we can measure the distance between two random variables quantile by quantile. After getting the well-defined measurement space, we employ equation (6) to solve for the time-invariant weights.

$$\min_{\lambda \in R^J} \int_0^1 \left| \sum_{j=2}^{j+1} \lambda_j F_{Y_{jt}}^{-1}(q) - F_{Y_{1t}}^{-1}(q) \right|^2 dq \quad s.t. \sum_{j=2}^{j+1} \lambda_j = 1 \tag{6}$$

Equation (6) illustrates how to find the weights to build the counterfactual distribution at time t. This optimization problem is a convex optimization, which means the local optimal must be the global optimal. Meanwhile, the method from equation (6) Follows the natural thought from the classical SCM. The next step constructs the numerical equivalence using monte carlo integration to solve the optimization problem defined in equation (6).

$$\min_{\lambda \in R^J} \frac{1}{M} \sum_{m=1}^{M} \left| \sum_{j=2}^{j+1} \lambda_j F_{Y_{jt}}^{-1}(V_m) - F_{Y_{1t}}^{-1}(V_m) \right|^2 \quad s.t. \sum_{j=2}^{j+1} \lambda_j = 1 \tag{7}$$

Equation (7) is the discrete equivalence of equation (6), where V_m is the m_{th} drawn from the uniform distribution with support from 0 to 1. By solving equation (7), we can get the optimal weights at time t. Unlike the method used in Gunsilius (2023), we randomly select

500 quantiles from a Beta(3, 3) distribution instead of from a uniform(0,1) distribution. This is because our first step involves performing quantile regression to estimate time-conditional empirical quantile values. If we were to select quantiles uniformly, observations at the extremes would receive disproportionately large weights. Consequently, regions with sparse observations would get an undesirably high number of estimates. We use quantiles drawn from a Beta(3, 3) distribution to better fit the data. However, the SCM framework requires time-invariant weights to build the counterfactual.

$$\vec{\lambda}^* = \sum_{t \leq T_0} \omega_t \vec{\lambda}_t^* \quad \text{for} \quad \omega_t \geq 0 \; and \; \sum_{t \leq T_0} \omega_t = 1 \tag{8}$$

Equation (8) offers a way to get the weights across time. The question is how to find the ω_t? In practice, $\omega_t = \frac{1}{T_0}$ Gunsilius (2023). Now we can use our optimal weights $\vec{\lambda}^*$ to build the counterfactual.

$$F^{-1}_{Y_{1t,N}} = \sum_{j=2}^{j+1} \lambda_j^* F^{-1}_{Y_{jt}} \quad \text{for } t > T_0. \tag{9}$$

From equation (9), we are able to get the counterfactual distribution using the counterfactual quantile function $F^{-1}_{Y_{1t,N}}$. When we get the counterfactual distribution, we can compare it with the distribution of the treated unit (province in China) after the treatment, where the difference will capture the GE treatment effect on the conditional distribution of corn yield in China at the province level.

Data

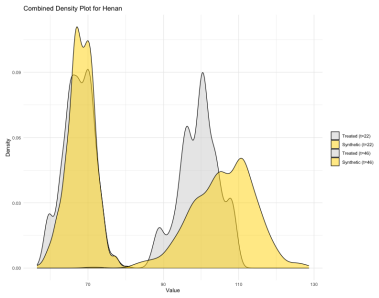
The corn yield data for China, the United States, Brazil, and Argentina all originate from public datasets. Specifically, the Brazilian data is sourced from The Brazilian Institute of Geography and Statistics; Argentina's data comes from National Public Administration Bodies, China's data is obtained from the National Bureau of Statistics, and the data for the United States is from Quick Stats in the USDA. The United States pioneered the adoption

of GE corn seeds as early as 1996. Argentina soon followed this in 1997, demonstrating a quick regional uptake. Brazil joined this trend in 2006, marking a significant step in the widespread acceptance of GE technologies in agriculture across these major economies.

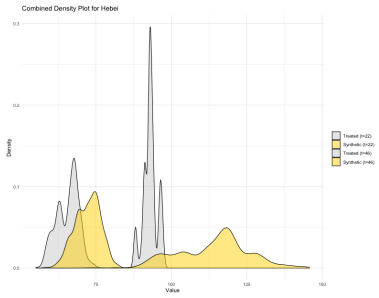
Results and Inference

Density Plots

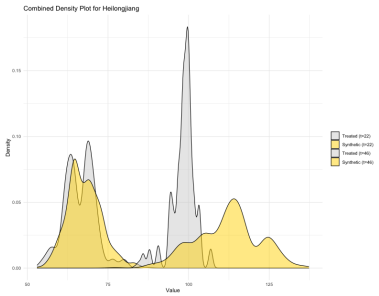
In this section, we will plot the conditional yield densities of 4 typical provinces in China, including the real densities and synthetic ones.



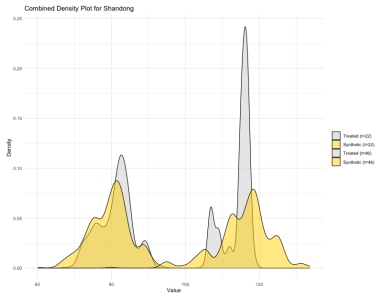
(a) Conditioanl densities of Henan



(b) Conditioanl densities of Hebei



(c) Conditioanl densities of Heilongjiang



(d) Conditioanl densities of Shandong

Figure 2: conditional yield density of corn in China

For each subgraph in Figure 2, the left figures show the status of fitting before treatment at time = T_{0-1}, i.e., the year 1995; the difference between the right part of densities demonstrates the GE effect on the yield density of corn in the year 2019.

Top Control Units

Table 1 shows the top 10 control units for Henan province in China. The DSC weights are not sparse because the time-invariant weights are the average weights across time. For a specific time, the optimal weights are sparse.

Provinces	**Country**	**Weights**
California	USA	0.1135
Mato Grosso	Brazil	0.1079
Salta	Argentina	0.1078
Chaco	Argentina	0.1014
Espírito Santo	Brazil	0.0850
Sergipe	Brazil	0.0711
Alagoas	Brazil	0.0697
Tucuman	Argentina	0.0650
South Dakota	USA	0.0530
La Pampa	Argentina	0.0360

Table 1: Top 10 Control Units Based on Average Weights Across Time

Permutation Test

We conducted a permutation test Abadie (2021) using the 2-Wasserstein distance for Henan Province in China. The major corn-producing provinces from Argentina, Brazil, and the United States were selected as units participating in the permutation test. They are:

- Argentina: Cordoba, Buenos Aires, Santa Fe, Santiago del Estero
- Brazil: Mato Grosso, Paraná, Goiás, Minas Gerais
- United States: South Dakota, Nebraska, Kansas, Minnesota, Iowa, Missouri, Wisconsin, Illinois, Indiana, Ohio

We find no treatment effects on the time-conditional distributions of corn yield in Henan province until 2019.

Jackknife Test

Due to the absence of an overall treatment effect on the time-conditional distribution of corn yield in Henan province until 2019, we next use the Jackknife to test for heterogeneous treatment effects across different quantiles. We mimic the process of bootstrapping test for the DSC estimator from Van Dijcke, Gunsilius, and Wright (2024) to test for heterogeneous treatment effects. Since Jackknife and bootstrapping belong to resampling methods, borrowing the spirit of the bootstrapping test from Van Dijcke, Gunsilius, and Wright (2024) is reasonable.

The Jackknife method is used to estimate the confidence interval of a parameter by systematically leaving out one observation at a time and recalculating the estimate. Here are the steps involved in calculating the Jackknife Confidence Interval for the treatment effects across the 500 quantiles:

1. **Compute the Jackknife Estimates** $\hat{X}_{(-i),j}$, where $i, j \in \{1, ..., 500\}$ and $i \neq j$:
   - $\hat{X}_{(-i),j}$ denotes the treatment effect at jth quantile calculated from leaving ith quantile out.
   - For each jackknife process, we use the DSC method to get the treatment effects across the rest of the 499 quantiles.
2. **Compute the Jackknife Mean** $\bar{X}_j$ **for each quantile**:
   - This is the average of jackknife estimates for the treatment effect at the jth quantile:
   $$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_{(-i),j}$$
   - $n = 500$ in our case
3. **Compute the Difference Between Each Jackknife Estimate and the Jackknife Mean**:

- The absolute difference at the jth quantile for the ith jackknife loop is computed as:

$$\text{Difference}_{i,j} = \left| \hat{X}_{(-i),j} - \bar{X}_j \right|$$

4. **Determine the Maximum Absolute Difference**:

   - The maximum absolute difference at the jth quantile is:

$$\text{MD}_j = \max(\{\text{Difference}_{i,j}\}_{i=1}^{500})$$

5. **Get the** $(1-\alpha)$ **Quantile value of** $\{MD_j\}_{j=1}^{500}$:

   - we use $t_{1-\alpha}$ to denote the $1-\alpha$ quantile value of $\{MD_j\}_{j=1}^{500}$.
   - α is the significance level. We use $\alpha = 0.05$.

6. **Construct the Confidence Interval**:

   - The confidence interval is given by:

$$\hat{T}_j \pm t_{1-\alpha}$$

   - $\hat{T}_j$ is the treatment effect for the jth quantile estimated by using the full data.

Figure 3 is a time panel showing the dynamic of the heterogeneous treatment effects and the confidence intervals calculated by the jackknife. In this graph, only the quantiles on the right half of the distribution show the treatment effects after nearly 13 years. The quantiles on the left half show no treatment effect even by 2019. Figure 4 compares the heterogeneous treatment effects at $t = 22(1995)$ with $t = 46(2019)$. 1995 is the last year during the pre-treatment period. Figures 3 and 4 mimic the bootstrapping graphs in Van Dijcke, Gunsilius, and Wright (2024).

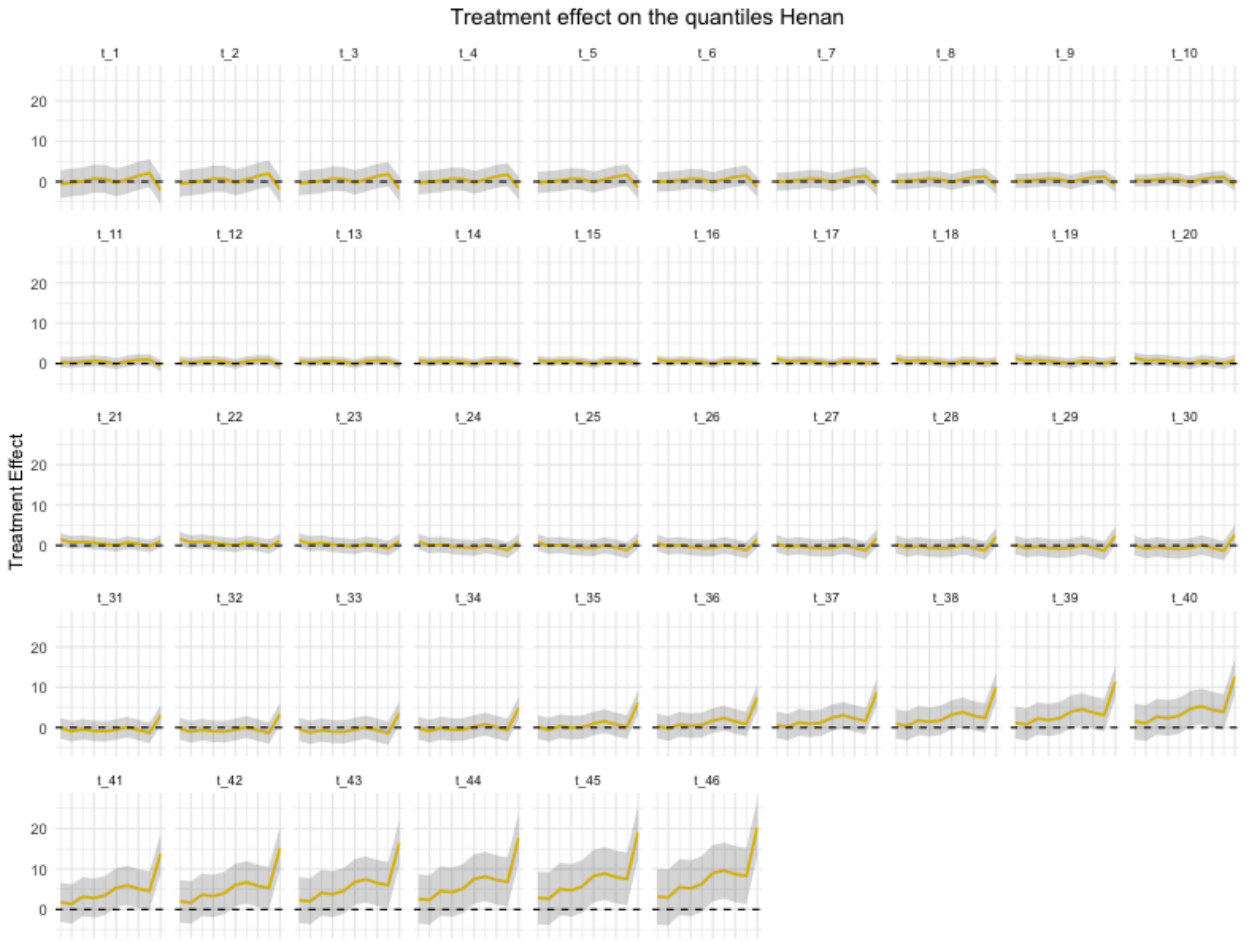
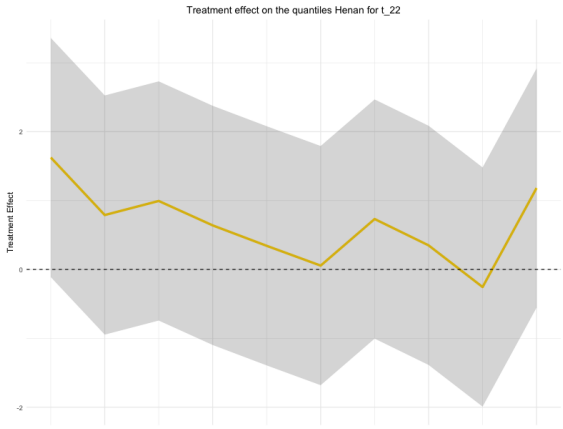
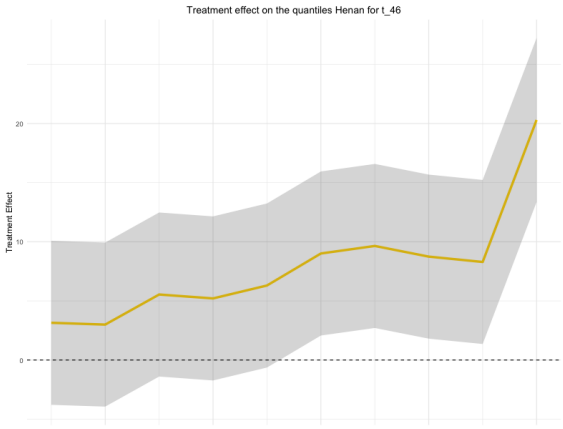


Figure 3: Dynamic treatment effects on the quantiles in Henan province



(a) Treatment effect on the quantiles in Henan at t_22



(b) Treatment effect on the quantiles in Henan at t_46

Figure 4: Comparison of Treatment Effects on Quantiles in Henan

References

Abadie, A. 2021. "Using synthetic controls: Feasibility, data requirements, and methodological aspects." *Journal of Economic Literature* 59:391–425.

Abadie, A., A. Diamond, and J. Hainmueller. 2007. "SYNTHETIC CONTROL METHODS FOR COMPARATIVE CASE STUDIES: ESTIMATING THE EFFECT OF CALIFORNIA'S TOBACCO CONTROL PROGRAM.", pp. .

Athey, S., and G.W. Imbens. 2017. "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic perspectives* 31:3–32.

Chen, Y.T. 2020. "A distributional synthetic control method for policy evaluation." *Journal of Applied Econometrics* 35:505–525.

FAO. 2016. "The state of food and agriculture: Climate change, agriculture and food security." Working paper, Food and Agriculture Organization of the United Nations, Rome.

Gunsilius, F.F. 2023. "Distributional synthetic controls." *Econometrica* 91:1105–1117.

Ker, A.P., T.N. Tolhurst, and Y. Liu. 2016. "Bayesian estimation of possibly similar yield densities: implications for rating crop insurance contracts." *American Journal of Agricultural Economics* 98:360–382.

Koenker, R., and K.F. Hallock. 2001. "Quantile regression." *Journal of economic perspectives* 15:143–156.

Lusk, J.L., J. Tack, and N.P. Hendricks. 2018. "Heterogeneous yield impacts from adoption of genetically engineered corn and the importance of controlling for weather." In *Agricultural productivity and producer behavior*. University of Chicago Press, pp. 11–39.

Muggeo, V.M., M. Sciandra, A. Tomasello, and S. Calvo. 2013. "Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology." *Environmental and ecological statistics* 20:519–531.

Scheitrum, D., K.A. Schaefer, and K. Nes. 2020. "Realized and potential global production effects from genetic engineering." *Food Policy* 93:101882.

Schuurman, D., and A. Ker. 2024. "Heterogeneity, climate change, and crop yield distributions: Solvency implications for publicly subsidized crop insurance programs." *American Journal of Agricultural Economics*, pp. .

Tolhurst, T.N., and A.P. Ker. 2015. "On technological change in crop yields." *American Journal of Agricultural Economics* 97:137–158.

Van Dijcke, D., F. Gunsilius, and A. Wright. 2024. "Return to Office and the Tenure Distribution." *arXiv preprint arXiv:2405.04352*, pp. .