To: **Damara Alves , Ministry of Women, Family and Human Rights, Brazil, and Ministry Data Scientists**

From: **Gaurish Katlana and Elton Vargas**

Date: **23rd April 2021**

Re: **Levelling Down: A Synthetic Control Method to Extend The Results of *"Soap operas and Fertility: Evidence from Brazil"* (2012)**

Executive Summary

"Soap Operas and Fertility: Evidence from Brazil" by Ferrara, Chong, and Duryea (2012) is a research paper that aimed to find the effect of soap operas on fertility in Brazil. We are interested in the effect of TV culture on family planning due to female characters communicating that freedom and career advancement can be achieved by having lower numbers of children. In the paper, the author expressed that people in the lower income strata display a greater reduction in fertility, suggesting that the effect of soap operas might not be the same for all cities as they differ in various cultural, economical and social aspects. As soap operas are nationalized, we propose that we can promote better family planning policies by calculating the local city level effect of the soap operas and thus, transmit soap operas to specific states as a low cost intervention to improve birth control in the country. We extend this study by comparing post-treatment birthrates in an AMC (Minimally Comparable Area) region that is in Cuiabá, a city in southern Brazil. We propose that we can better estimate the effect of Globo coverage on fertility decline using a Synthetic Control Method for this region because it evens out time-dependent variables, local area effects, and civil unrest in the country (that is unaccounted for in the original study). The original study identifies soap operas' effect at the national level without accounting for regional differences. We found that economic and education variables display low standard errors signifying that they are effective at accounting for regional differences. Our synthetic control model found that the effect of Globo soap operas on fertility rates in Cuiabá is approximately -12%, suggesting that the negative effect of soap operas might have been exacerbated in low-income and undereducated areas. Thus, we recommend that the effect on behaviors of content transmitted in popular media, such as soap operas, is further investigated and leveraged to conduct low-cost high-impact public policies.

Introduction

This brief memo summarizes the methodology and results of a Synthetic control Method that we used to estimate the treatment effect on Brazilian fertility rates in the city of Cuiabá as an extension to the paper "Soap Operas and Fertility: Evidence from Brazil" by Ferrara, Chong, and Duryea (2012). We hypothesize that Synthetic Control Matching can successfully achieve a very high degree of pretreatment match, while being robust at exogenous effects, such as the unconfounded effect civil unrest in several parts of the country may have had on fertility rates.

By creating an artificial control unit (our synthetic control result) for a specific county, we assure that all pretreatment conditions are appropriately accounted for in that specific unit. Our high balance in pre-treatment covariates suggests that the use of Synthetic Control might be robust in calculating the true effect of Globo soap operas on the city's fertility rates. Using this method, we can create a Synthetic Control unit for every county. In this paper, we prepare a Synthetic Cuiabá county, which we expected that it would reveal a significant negative effect of soap operas on fertility rates. The main advantage of this method is that we can see the effects of Globo entry on individual counties, and thus, provide more precise guidance after assessing how Globo soap operas' transmission specifically impacted fertility on a given county. However, the counterargument to that advantage is that the data at hand only provides 7 years of preliminary data meaning that our Synthetic Control units might not have enough pre-treatment data to infer the true treatment effect.

Though the original paper calculated the soap opera effect on fertility rates using linear regression and controlling for covariates correlated with Globo network entry, we argue that the regression did not provide an effective local model as we can see that the variance across various counties is large and thus, the regression predicts poorly county level effects. We believe that county-level policies could benefit more from the more coarse-grained information about the effect of Globo entry on the county's fertility rates. We also believe that the Synthetic Control unit will better fit the data at the local level. Given the large number of counties (approximately 360 counties) that are available in the dataset, the Synthetic Control method would give a more robust and accurate estimate of both the global and the local soap opera effect.

1. Data and modifications to the data

We summarize the pre processing of data below (See the implementation required in Appendix A: **R** code).

The original data set is a combination of filtered census data from 1979 to 1991, combined with Globo coverage data obtained through geographic matching. The original dataset had 300,250 observations and 36 columns (you can find the meaning of variables in Appendix C):

```
[1] "id"             "rural"          "electricity"   "tv"
[5] "weight"     "catholic"       "yrs_edu"


   "married"
[9] "employed" "yrsedu_head" "uf_code      " "age          "
[13] "year"      "B            "      "stock        " "globocoverage1"
[17] "yrsexp       " "yrsexp1019    " "yrsexp2029    " "yrsexp3039    "
[21] "yrsexp4049    " "geoarea80     " "ipc_gdp      " "ipc_renta     "
[25] "wealth_noTV   " "Doctors      " "agesq        " "stocksq      "
[29] "area         " "yr1stcov     " "Blag         " "globocov1lead "
```

[33] "cov1wealth " "cov1eduhd " "cov1edu " "unit "

The data set is large however compared to the female population of brazil (61.8 million in 1981) that it is sampled from the dataset is still small.

Step 1
We removed all outliers in county size because larger counties tend to be sparsely populated with fewer rows in the data. Hence, we dropped all counties with an area greater than 5994 km^2. We convert all columns into numeric data type and convert amc_code (representing the county name) to county IDs.

Step 2
Because we are looking at counties rather than individual, we converted all the columns of the data from individual to county level. We separate the data based on two factors amc_code and year. We dropped all counties with less than 1000 data entries. We then converted all the data entries into a single entry representing the mean for every county at every year available in the dataset. The usage of mean is justified here because there are no outliers in any columns (defined as having more than 3 times the standard deviations) and the condensed mean data obtained is representative of the county and the year it is taken from.

Step 3
We also dropped six time variables created in order to facilitate the regression analysis on the data; those columns are yrsexp,yrsexp1019,yrsexp2029,yrsexp3039,yrsexp4049, and blag.

2. Extension 1: Synthetic control procedure
The Synthetic control procedure involves using a combination of various counties from the control group and assigning each county a weight to better resemble the real county's pre Globo-entry characteristics. The sum of all the used counties' weights should add up to one. In this way, the Synthetic county is a linear combination of counties that represents Cuiabá (our chosen treatment unit) as closely as possible before Globo entry (from 1979 to 1986), which is the previous period to our treatment (exposure to soap operas). After the pretreatment period, we compare the results between the Synthetic county and the treated county. The difference in outcomes between the two counties is the treatment effect as we are effectively comparing the potential outcomes for the same unit, whether with real or synthetic data.

The given county received the treatment that is Globo coverage in 1986. However, because it would take some time to see the effect of soap operas on the fertility rate (e.g., 1-year duration of a soap opera), we consider the treatment to be given in 1987.

We specified the following covariates to be considered for the synthetic function

"Rural" ,"electricity", "tv" ,"weight" ,"catholic" , "yrs_edu" , "married" , "employed" ,"yrsedu_head" ,"uf_code" , "age"

The covariates are selected to account for maximum information and R^2values from the Appendix B table 1.

We looked at the balance between the covariates of the synthetic unit and treated unit before the treatment (year = 1987) to find the dimensions of the data.

We then proceed to conduct sensitivity analysis test on the obtained results to find:

1) sensitivity of the fit between the synthetic state and the treated state outcomes in the pretreatment period
2) sensitivity of the synthetic state outcome in the treatment period

3. Results

After converting the data and data preprocessing we obtained the following results from the Synthetic control method, we can see from the figure below the fertility rate for the synthetic *Cuiabá* unit and the real *Cuiabá county*. Ideally, the two lines should follow each other closely before treatment, we will quantify the fit using root mean square prediction error.

The RMSPE for *Cuiabá* is 0.0261997 and that for the synthetic region is 0.01737728 which is very small, indicating that the synthetic control method has achieved a high degree of pretreatment matching.

Figure 1: Pretreatment Outcome Gap between Real Cuiabá and its Synthetic Unit

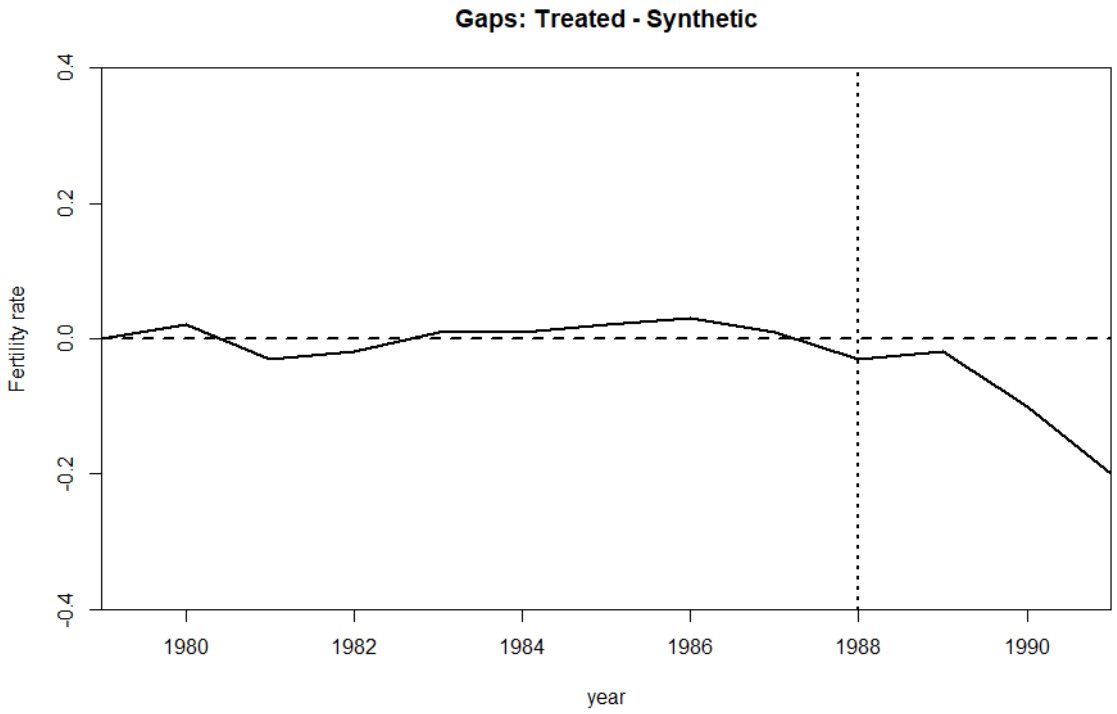

Figure 2: The difference of the outcome variable between treated and Synthetic control units.

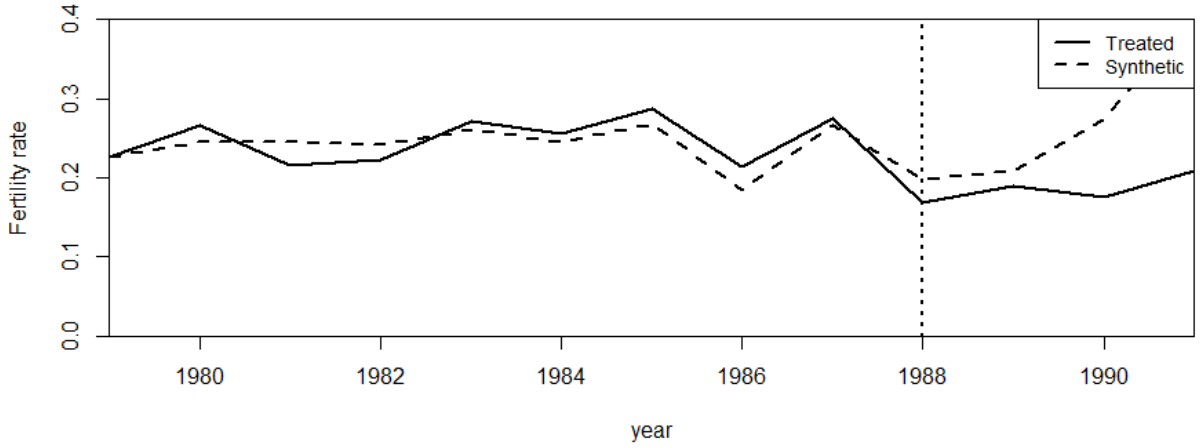

Table 1: Mean covariates results for the Synthetic control unit, treated unit and the data mean

Variable	Treated	Synthetic	Sample mean
rural	0.640	0.691	0.558
electricity	0.293	0.299	0.310
tv	0.827	0.838	0.853
weight	6.891	6.990	7.454
catholic	0.867	0.906	0.907
yrs_edu	2.053	2.046	2.396
married	0.653	0.643	0.651
employed	0.267	0.223	0.208
yrsedu_head	1.107	1.588	1.888
uf_code	19.000	19.100	26.321
age	27.680	27.683	28.232

Table 2: Weight of various variables

Vairble name	Weight of the variable
rural	0.011
electricity	0.328
tv	0.118
weight	0.006
catholic	0.004
yrs_edu	0.16
married	0.098
employed	0.021
yrsedu_head	0.01
uf_code	0.048
age	0.196

Table 3: The composition of the synthetic unit

Unit Number	Weight
-2139698080	0.235
-1619397197	0.001
-830359686	0.001
-705374624	0.002
-467743082	0.095
45369086	0.132
93798556	0.093
388384962	0.367
1886063827	0.074

Table 4: Loss Function of the Synthetic Control Model

Loss W	Loss V
0.00261997	0.001737728

Our estimate of the effect of Globo coverage in *Cuiabá* is given by the difference between the fertility rate right after 1987, or in the year 1988. We can see that the lines start diverging after 1988, which displays an effect of -0.12, meaning that our model found that Globo soap operas entry to Cuiabá caused a decline of -12% in the county's fertility rate.

This result is a good approximation of the true effect Globo soap operas on Cuiabá fertility rates because our model is robust and well balanced. The greatest difference between the mean of a treatment and control covariate was 0.4 in the *yrsedu_head* variable. Other than that, there are small differences in means of about 0.1, evidencing that our model achieved high balance in the treatment and control units' covariates.

Covariates *electricity, age* and *yrs_edu* received the most weight in our synthetic control unit, suggesting that age, education level, and access to electricity affect fertility rates heavily. Moreover, since we dealt with a data set of 2.4 million units, it was impossible for us to track down the name of the county that received the most weight in our synthetic control, however, we provided unit IDs, which can identify which units received the most weight in our synthetic control unit if that's desirable. However, fairly distributed unit weights (less than 0.3) may indicate that our model isn't overdependent and thus, highly sensitive to a single unit.

4. Policy implications of our findings

With our Synthetic Control analysis for Cuiabá, we found that the effect of soap operas transmission on fertility rates was magnified (-12% effect). Thus, computing a Synthetic Control unit for each county may help understand better the effect of soap operas messages on fertility rate at a local more granular level. The differences in soap operas' effects among counties is particularly relevant to avoid spending ineffective monetary resources on nation-wide telecommunication interventions that may only affect a couple of counties. Given the characteristics of Cuiabá, a potential implication of our findings might be that soap operas messages are more impactful at low-income and undereducated areas. However, this supposition needs to be corroborated with synthetic control models for each county. With these specific results, we suggest that the Ministry: 1) takes into account county differences in public policy programs' effects, and 2) further investigates and leverages how might messages sent in

popular content, such as soap operas, might lead to improvements in population behavior (e.g., decrease in child domestic violence), particularly in low-income communities.

5. Recommended additional analysis

We recommend extending this proposed case study to include an Augmented Synthetic Control method, which is used when pretreatment fit is infeasible. We achieved a high degree of pretreatment fit; however, this optimized method will be a strong way of confirming our results. An additional advantage is that this method uses an outcome bias correction algorithm to reduce the bias from the results of the Synthetic Control method, which will lead to improvements in our effect approximation.

We also recommend using an “in-time placebo” test, where we can reassign the treatment to occur at 1983 (mid-way between 1980-1986 pre-treatment period), and see if our synthetic control still resembles the trajectory of our treatment unit to assess the robustness of our model.

6. Conclusion

To extend the results of *“Soap Operas and Fertility: Evidence from Brazil”,* we generated a Synthetic Control unit for the Cuiabá county that allowed us to assess the effect of soap operas on fertility rates at a local level. Our Synthetic Control achieved high balance between treatment and control covariates, suggesting that our model yields a fairly good effect estimate. We found that the effect is about -12%, meaning that fertility rates in the county significantly declined after soap operas’ TV debut in Cuiabá. Since Cuiabá is a low-income county, we recommended that the Ministry of Women and Family uses more often these popular media contents in these locations to encourage positive behavior in the Brazilian population (e.g., ending child domestic violence). To improve our findings, we suggest: 1) the computation of a Synthetic Control unit for each county to compare and contrast the local effects, 2) the use of the Augmented Synthetic Control method to improve our model’s fit and reduce its outcome bias, and 3) an “in-time placebo” test to evaluate the robustness of our model.

APPENDIX A: R code

```
```{r}
rm(list = ls())
#Installing libraries that may or may not be used
library("haven")
library("foreign")
library("corrplot")
library("cobalt")
library("MatchIt")
library("Matching")
library("gridExtra")
library("ggplot2")
library("rbounds")
library("rgenoud")
library("Synth")
library("dplyr")
library("digest")
install.packages("devtools")
library("devtools")
devtools::install_github("ebenmichael/augsynth")
install.packages("augsynth")
library("augsynth")
# we need to set the seed of R's random number generator,
# in order to produce comparable results
set.seed(123)


```

```{r}
#The dataset can be downloaded from https://tinyurl.com/CS112Final
Indiv <- read_dta("Indiv.dta")
Indiv <- Indiv[Indiv$geoarea80>=5994,]
#Removing non numeric columns and converting county name column to a
numeric column
Indiv$unit <- c(unlist(lapply(Indiv$amc_code, digest2int)))
Indiv <- Indiv %>% select_if(is.numeric)

#Aggregate function to create a panel data frame for
Synthetic_method <- data.frame(Indiv)
Synthetic_method <- Synthetic_method[0,]
```
```

```
Synthetic_method <- aggregate(Indiv, list(Indiv$unit,Indiv$year), mean)

# We will look at country AMC 15 0110 where the treatment was given in
treat_year
treat = 1499611235
treat_year = 1988

Synthetic_method <- as.data.frame(Synthetic_method)
#Drop observations that received the globo coverage before treat_year

#Counties that received television before treat_year
AMCs <- c(unique(Synthetic_method[Synthetic_method$globocoverage1==0 &
Synthetic_method$year == treat_year,]$unit),treat)
#Filtering the data
Synthetic_method <- Synthetic_method %>%
  filter(unit %in% AMCs)

#Dropping NAs
Synthetic_method <- subset(Synthetic_method, select
=-c(yrsexp,yrsexp1019,yrsexp2029,yrsexp3039,yrsexp4049,Blag))
#Using genetic matching in order to prematch data to reduce computation
time in SCM
Synthetic_method$Treated <- Synthetic_method$unit == treat &
Synthetic_method$year >treat_year

#Preparing list of controls and treated units
J <- unique(Synthetic_method$unit)
J = J[J!=treat]
```

```{r}
#Extension to the study

#Extension 1, Using a synthethic control to account for fertility
difference given entry of rede globe in
# 1991 in county X. We will control a synthetic county X using counties
that didn't have redeglobe entry in any

dataprep.out <- dataprep(
  #data
  foo = Synthetic_method,
  #predictors
```

```
  predictors = c("rural"          ,"electricity",
"tv"             ,"weight"         ,"catholic"       ,"yrs_edu"
,"married",
"employed"       ,"yrsedu_head"    ,"uf_code"        ,"age"),
  predictors.op = 'mean', # the operator
  special.predictors = NULL, #We don't have special values
  #The outcome variable
  dependent = 'B',
  #identifiers
  unit.variable = 'unit', #Numeric column with unit numbers
  time.variable = 'year', #Column with time

  treatment.identifier = treat,#the treated case
  controls.identifier = J,#the control cases; all others #except number 17

  time.predictors.prior = c(1979:1979), #Averaging outcomes before

  time.optimize.ssr = c(1979:treat_year),#the time-period over which to
optimize
  time.plot = c(1979:1990)

  )


#######################################################
#############Running synthetic control##############
#######################################################

synth.out <- synth(data.prep.obj = dataprep.out) #default optimx function
is c("Nelder-Mead", "BFGS").

weights <- synth.out$solution.v

#Plotting results
synth.tables <- synth.tab(dataprep.res = dataprep.out,synth.res =
synth.out)
print(synth.tables)

gaps.plot(synth.res = synth.out,dataprep.res = dataprep.out,Ylab =
c("Difference in outcome"),Xlab = c("Time"),Main = c("Gaps: Treated -
Synthetic"),tr.intake = treat_year,Ylim = NA,Z.plot = FALSE)
```

```
path.plot(synth.res = synth.out, dataprep.res = dataprep.out,
          Ylab = "Fertility rate", Xlab = "year",
          Ylim = c(0, 0.4), Legend = c("Treated County",
                                       "synthetic county"),
Legend.position = "topright")


#Extension 2 using augmented solution

syn <- augsynth(B ~ Treated, unit, treat_year, Synthetic_method,
                progfunc = "None", scm = T)
summary(syn)
plot(syn)

```

```
#Replication of table 1
lm01 <- lm(B~globocoverage1 + as.factor(Indiv$year) -1 ,
data=Indiv,weights= Indiv$weight)
summary(lm01)


lm02 <- lm(B~globocoverage1 +
as.factor(Indiv$year)+as.factor(Indiv$uf_code) +1 , data=Indiv,weights=
Indiv$weight)
summary(lm02)

lm03 <- lm(B~globocoverage1 +
as.factor(Indiv$year)+as.factor(Indiv$amc_code) -1 , data=Indiv,weights=
Indiv$weight)
summary(lm03)

lm04 =
lm(B~globocoverage1+married+yrsedu_head+wealth_noTV+catholic+rural+Doctors+
ipc_renta+age+agesq+stock+stocksq+as.factor(year),data=Indiv)
summary(lm04)

#cols. 4-6
Col4 <- lm(B ~ globocoverage1 + married + yrsedu_head + wealth_noTV + catholic +
rural +Doctors + ipc_renta + age + agesq + stock + stocksq + i.year , w=weight,
cluster(amc_code))
```

```
Col5 <- glm(B ~ globocoverage1 + married + yrsedu_head + wealth_noTV +catholic +
rural + Doctors + ipc_renta+age+agesq+stock+stocksq+i.year+as.factor(uf_code),
w=weight,cluster(amc_code))

Col6 <- glm(B~ globocoverage1 + married + yrsedu_head + wealth_noTV + catholic +
rural + Doctors + ipc_renta + age + agesq + stock + stocksq + i.year +
as.factor(amc_code)  ,w=weight, cluster(amc_code) )
```

Appendix B: Replication of Table 2 - Globo Coverage and Fertility

The table is made using the code in the appendix A, in this section some of the variables could not be replicated and such are bolded.

Panel A data

	[1]	[2]	[3]
Globo coverage	-0.0269	-0.0115	**-0.07**
Constant	0.1177	0.1126	**0.158**
R^2	0.003	0.006	**0.002**

Panel B data (Coefficients omitted for clarity) (see code for more info)

	[4]	[5]	[6]
Globo coverage	-0.0075	-0.0115	-0.0047
Constant	-0.193	0.1126	-0.1877
R^2	0.046	0.006	0.050
Age	0.023	0.023	**0.023**
Stock of children	0.0029	**0.0028**	0.0017
Education of head	-0.0002	-0.0002	-0.0002
Wealth	-0.0208	-0.0204	-0.0205

Appendix C: Column names and meanings

"id" : Identifying number for a female
"rural" : 1 if living in rural area; 0 otherwise
"Electricity" : 1 if there is access to electricity; 0 otherwise
"Tv" : 1 if has access to TV; 0 otherwise
"weight" :
"catholic" : 1 if has catholic; 0 otherwise
"yrs_edu" : Years spent in school from 1st grade
"Married" : 1 if married; 0 otherwise
"employed" : 1 if employed; 0 otherwise
"yrsedu_head" : Number of years studied by parnts
"uf_code " State code
"age "
"year "
"B ": If the person gave birth in the current year
"stock " : Number of children
"Globocoverage1": Whether the globo network was available in the area 1 year prior
"yearsxp" "yrsexp1019 " "yrsexp2029 " "yrsexp3039 ""yrsexp4049 ": Years exposed to globonetwork between age 10-19, 20-29, 30-39, 40-49. (Dummy time variables)
"geoarea80 " : Area of the AMC
"ipc_gdp " : GDP per capita of the regio
"ipc_renta " : Average rent in the area
"wealth_noTV " : Television set per household
"Doctors " : Healthcare workers per 1000 people
"agesq ": Age squared / 100
"stocksq ": stock squared / 100
"area " numeric value for amc code
"yr1stcov " :Year when the globo signal entered AMC
"Blag ": Adopted children
"cov1wealth ": Standard deviations away from income
"cov1eduhd " : Years in education before coverage
"cov1edu ": Years in education after coverage

References

Ben-Michael, E., Feller, A., & Rothstein, J. (2018). *The Augmented Synthetic Control Method*. ArXiv.org. https://arxiv.org/abs/1811.04170

Ferrara, E. L., Chong, A., & Duryea, S. (2012). Soap Operas and Fertility: Evidence from Brazil. *American Economic Journal: Applied Economics*, *4*(4), 1–31. https://doi.org/10.1257/app.4.4.1

Eliana La Ferrara, Chong, A., & Duryea, S. (2019). Replication data for: Soap Operas and Fertility: Evidence from Brazil. *Openicpsr.org*. https://doi.org/10.3886/E113835V1

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746