README

Overview

The code in this replication package allows to reproduce all the results reported in the paper "Losing control (group)? The Machine Learning Control Method for counterfactual forecasting". This folder contains the following subfolders:

- Data: `.csv` files with the data to replicate the empirical analysis
- Functions: `.R` functions used for the analysis and for simulations
- Codes: `.R` files with the R codes to replicate the analysis (file "code.R") and simulations ("sim_code.R")
- Output: the exported tables in `.csv` format and figures in `.png` format resulting from executing "code.R" and "sim_code.R". [TO DO]

The paper introduces a novel methodology to estimate causal effects on panel data (i.e., multiple units observed in time) in the presence of an intervention/policy introduction by contrasting the observed outcome post-intervention with a predicted counterfactual based on a data-driven Machine Learning approach. The methodology is tested in a simulation study and then used to evaluate the causal effect of [COVID on Invalsi – CHANGE THIS, SAY BETTER].

Data Availability and Provenance Statements

The data used for the empirical analysis have been obtained upon request made to the data provider (i.e., the chief marketing officer of the supermarket chain). The data provider granted permission to share the data for scientific purposes in an anonymized fashion (i.e., without reference to the product). Under this agreement, the authors have made the data publicly available in the supplementary material of the manuscript (see the "Summary of Availability" statement below).

Statement about Rights

☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
☑ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Summary of Availability

The data for this project are confidential and the data provider wish to remain anonymous. However, upon permission granted by the data provider (see the "Statement about Rights" above) the authors of the manuscript deposited the data at Oxford University Press in the supplementary material of the paper.

The authors therefore confirm that all the data are now publicly available and can be used for replication purposes.

☑ All data **are** publicly available.

Dataset list

The "Data" folder contains the following files:

- `dates.csv`: vector of dates, from September 1, 2017 to April 30, 2019.
- `dummies.csv`: pre-generated dummies for Italian holidays and day-of-the-week effects
- `prices.csv`: daily prices of all goods included in the analysis
- `sales.csv`: daily number of units sold of each product

Description of programs

The "Functions" folder contains the following files:

- `Setup.R`: function and code to set up the environment
- `Functions.R`: main functions for estimating the causal effect
- `Plot_Functions.R`: plotting functions
- `Sim_Functions.R`: functions used in the simulation study

Further details about the input/output of each documented function are reported in the `.R` files.

Replication Instructions

Open the file "code.R", update your directory and execute the code. The code runs independently. For the simulation study, open the file "sim_code.R", update your directory and execute the code. The code runs independently. The latest version of R can be download from `https://www.r-project.org`. All the outputs (tables in `.csv` format and figures in `.png` format) are automatically exported and saved in the "Output" folder. For a complete list of the exported tables and figures see the section "List of tables and programs" below.

Computational requirements

List of R packages that are needed in order to replicate the results:

- R 4.2.1
  - `tseries` (0.10.50)
  - `forecast` (8.16)
  - `ggplot2` (3.3.5)

- pracma (2.3.8)
- ltsa (1.4.6)
- stats (4.2.1)
- gridExtra (2.3)
- parallel (4.2.1)
- the file “Setup.R” will install all dependencies (latest version). If, for any reason, “Setup.R” fails to install some or all the required packages, package installation should be done manually by executing the following line of code `install.packages("package name")` in the R console .

The results presented in the paper have been obtained under R version 4.2.1 on Windows 10 Home 21H1 Intel(R) Core(TM) i7-8550U CPU 1.80GHz 1.99 GHz. The approximate time needed to reproduce the analyses on a standard (2022) laptop machine is 10 mins. As for the simulation study, we experienced approximately 9 hours on our machine by employing code parallelization on 7 logical cores. If your machine uses less than 7 cores, computational time may be considerably higher. To reduce the computational time of simulations, we also used a server platform running under Ubuntu 22.04 LTS. The simulations' results reported in the paper and in the online supplement have been obtained on this platform under R version 4.2.1.

List of tables and programs

The provided codes reproduce selected figures and tables in the main paper and in the supplement. For each of them, we specify below the line of the program that generates them and the name of the corresponding file in the “Output” folder. Notice that all the tables are exported and saved in `.csv` format and, as such, they can be opened on many applications (e.g., Microsoft Excel, OpenOffice Calc, Notepad, Google Sheet).

The figures and tables reported in the main paper are reproduced as follows.

Figure/Table #	Program	Line Number	Output file	Note
Table 1	sim_code.R	215	table1.csv	Same as Table 2 in the supplement
Table 2	code.R	137	table2.csv	
Table 3	code.R	138	table3.csv	
Figure 1	sim_code.R	149	figure1.png	Same as Figure 2 in the supplement
Figure 3	code.R	120	figure3.png	
Figure 4	code.R	125	figure4.png	

The figures and tables reported in the online supplement are reproduced as follows.

Figure/Table #	Program	Line Number	Output file	Note
Table 1 S	sim_code.R	212	table1S.csv	
Table 4 S	sim_code.R	218	table4S.csv	
Table 5 S	sim_code.R	221	table5S.csv	
Table 6 S	code.R	140	table6S.csv	
Table 8 S	code.R	140	table8S.csv	
Figure 1 S	sim_code.R	179	figure1S.png	
Figure 3 S	sim_code.R	189	figure3S.png	

Please note that Tables 1S, 4S, 5S in the supplement and Table 1 in the main paper reproduce simulation results under different scenarios, namely, fictional interventions yielding to an increase or a decrease in the outcome level. Among them, we tested a scenario where the outcome level increases by 100%; since the results obtained are not particularly meaningful, they are not reported in the manuscript but they are reproduced here (see the last row of the aforementioned tables); this is done for transparency reasons and for the benefit of the interested readers.

References

Anonymous firm. (2019). “Daily sales data of selected brands for the period September 2017-April 2019” Unpublished data. Accessed April 30, 2019.