Assessing the performance of nuclear norm-based matrix completion methods on CO_2 emissions data

Rodolfo Metulini[1] Francesco Biancalani [2] Giorgio Gnecco [2] Massimo Riccaboni [2]

[1]Department of Economics, University of Bergamo

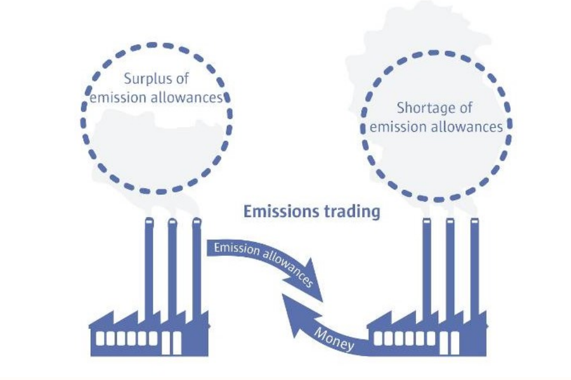
[2]Laboratory for the Analysis of CompleX Economics Systems (AXES), IMT School for Advanced Studies Lucca

Framework

Carbon Dioxide (CO_2) emissions represent a rising concern in relation to pollution and climate change (Yoro & Daramola, 2020)

Economic systems produce large amounts of CO_2 by the use of fossil energy. Governments are addressing the production to new systems aimed to minimize emissions.

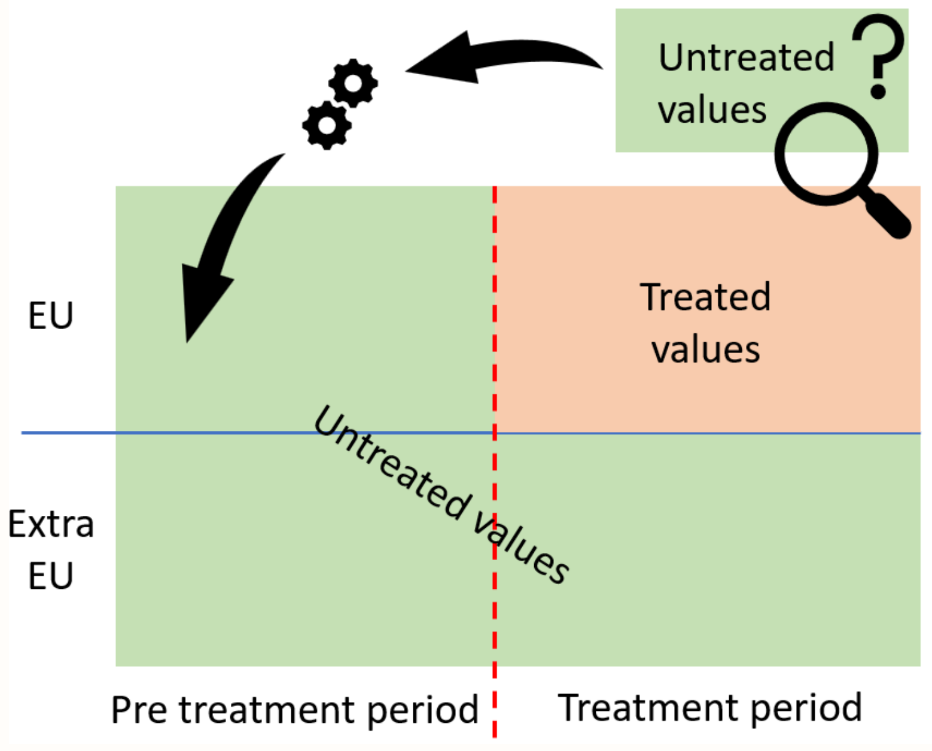
The European Union (EU) implemented a market of emission rights called the **Emissions Trading System** (ETS) that was launched in 2005, aimed at reducing greenhouse gas emissions.


Surplus of emission allowances
Shortage of emission allowances
Emissions trading

A **counterfactual analysis** for policy evaluation would permits to quantify the reduction of CO_2 emissions due to the ETS

Aim

Due to the ETS policy, untreated CO_2 emissions are **unknown** for the EU countries (treated) in treated years. **Matrix Completion** (MC) (Hastie et al., 2015) is a supervised statistical learning method to reconstruct a partially incomplete matrix.



We use MC to generate estimates of such untreated CO_2 emissions based on values of the EU countries in the pre-treatment period and on values of extra-EU countries in the treatment period.

To obtain a **robust** counterfactual, we have to study the performance of MC method in reconstructing the original matrix (in absence of treatment). We develop a simulation study to test the **performance** of Nuclear Norm-based MC methods for panel data.

Methodology

Given a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, MC works by finding a suitable low-rank approximation of $\mathbf{M}$, by assuming the model $\mathbf{M} = \mathbf{C}\mathbf{G}^\top + \mathbf{E}$, where $\mathbf{C} \in \mathbb{R}^{m\times r}$, $\mathbf{G} \in \mathbb{R}^{n\times r}$, whereas $\mathbf{E} \in \mathbb{R}^{m\times n}$ is a matrix of errors. Mazumder (2010) optimization problem - MC Baseline ($\mathtt{MCB}$):

$$\underset{\widehat{\mathbf{M}}\in\mathbb{R}^{m\times n}}{\text{minimize}} \quad \left(\frac{1}{|\Omega^{\mathrm{tr}}|}\sum_{(i,j)\in\Omega^{\mathrm{tr}}}\left(M_{i,j}-\widehat{M}_{i,j}\right)^2+\lambda\|\widehat{\mathbf{M}}\|_*\right)$$

Athey et al. (2021) methodological advancements (MC Fixed Effects - ($\mathtt{MCFE}$) and MC Time Fixed Effects - ($\mathtt{MCTFE}$)) explicitly include individual and time fixed effects in the optimization problem:

$$\underset{\widehat{\mathbf{L}}\in\mathbb{R}^{m\times n},\widehat{\Gamma}\in\mathbb{R}^{m\times 1},\widehat{\Delta}\in\mathbb{R}^{n\times 1}}{\text{minimize}} \quad \left(\frac{1}{|\Omega^{\mathrm{tr}}|}\sum_{(i,j)\in\Omega^{\mathrm{tr}}}\left(M_{i,j}-\widehat{M}_{i,j}\right)^2+\lambda\|\widehat{\mathbf{L}}\|_*\right)$$

$$\text{subject to} \quad \widehat{\mathbf{M}} = \widehat{\mathbf{L}} + \widehat{\Gamma}\mathbf{1}_{\boldsymbol{n}}^\top + \mathbf{1}_{\boldsymbol{m}}\widehat{\Delta}^\top$$

$\widehat{\Gamma}\mathbf{1}_{\boldsymbol{n}}^\top$ and $\mathbf{1}_{\boldsymbol{m}}\widehat{\Delta}^\top$ model row (individual) and column (time) fixed effects.
The nuclear norm $\|\widehat{\mathbf{L}}\|_*$ is used instead of $\|\widehat{\mathbf{M}}\|_*$, differently from $\mathtt{MCB}$.

References

[1] Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K.: Matrix completion methods for causal panel data models. Journal of the American Statistical Association **116(536)**, 1716-1730 (2021)

[2] Corsatea T.D., Lindner S., Arto, I., Roman, M.V., Rueda-Cantuche J.M., Velazquez Afonso A., Amores A.F., Neuwahl F.: World Input-Output Database Environmental Accounts. Update 2000-2016, EUR 29727 EN, Publications Office of the European Union, Luxembourg (2019)

[3] Hastie T, Tibshirani R, Wainwright M, Statistical Learning with Sparsity: The Lasso and its Generalizations. CRC Press, New York (2015)

[4] Mazumder R, Hastie T, Tibshirani R,: Spectral Regularization Algorithms for Learning Large Incomplete Matrices. Journal of Machine Learning Research **11**, 2287-2322 (2010)

[5] Yoro, K. O., & Daramola, M. O., CO2 emission sources, greenhouse gases, and the global warming effect. In: Advances in carbon capture, pp. 3-28. Woodhead Publishing (2020)

Simulation Study

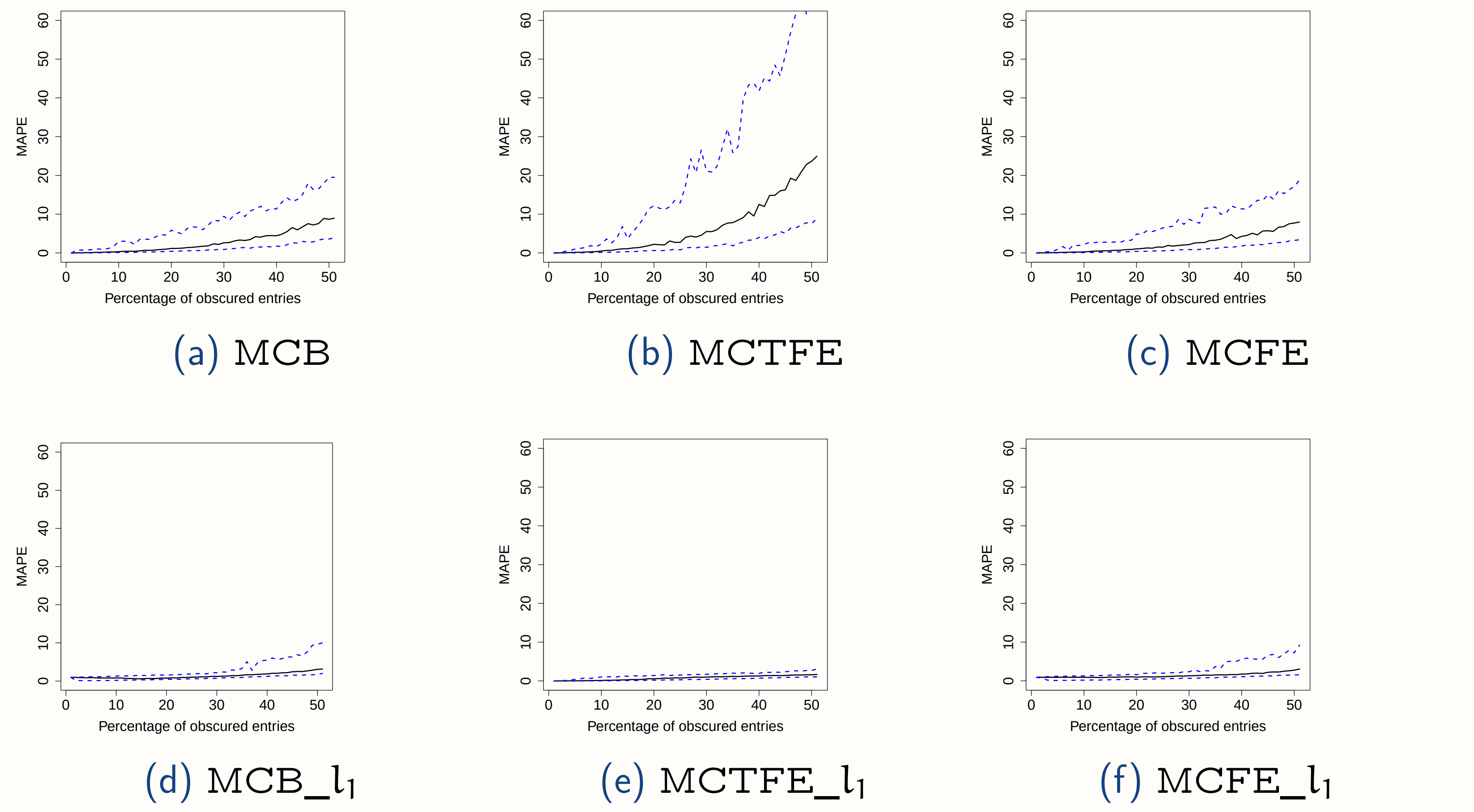
Free database on total CO_2 emissions (in thousand of tons) by country and sector (Corsatea et al, 2019), covering years 2000 – 2016 and 42 countries (29 EU + 13 extra-EU).

Years: from 2000 to 2005, to avoid treatment effects due to ETS. **Countries**: 26 (14 EU + 12 extra-EU, dropped small and extra-EU countries with EU agreements).

We compare the performance of $\mathtt{MCB}$, $\mathtt{MCTFE}$ and $\mathtt{MCFE}$, with respect to the **original matrix** and to a l_1 row-normalization by country, using **Root Mean Square Error** (RMSE) and **Between Deviance Percentage Ratio** (BDPR).

Unknown entries from 0 to 50%. **200 replications**, where the missing entries (test set) are chosen at random according to the desired percentage.

Computations performed with `mcnnm_cv` function in `MCPanel` R package.

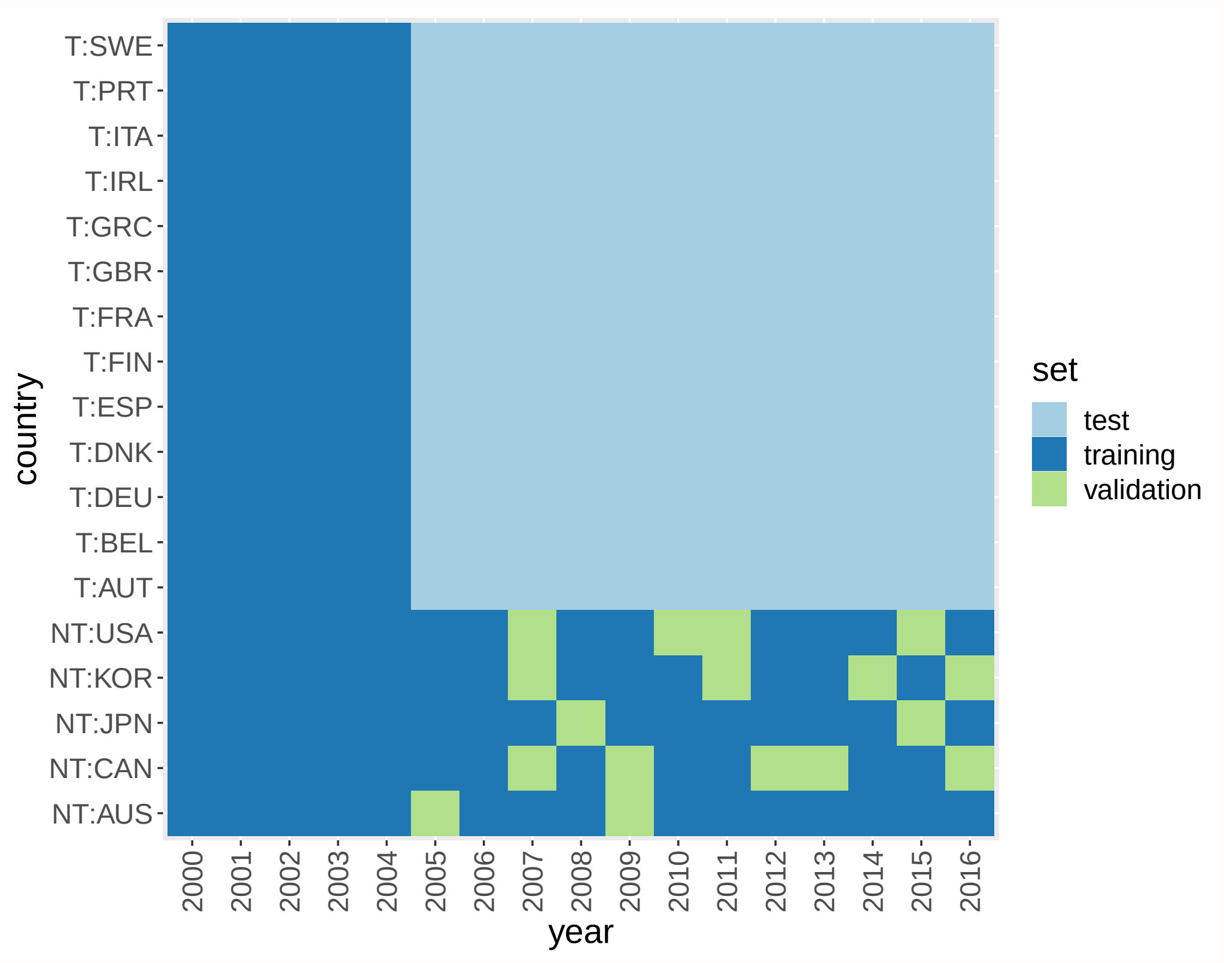


(a) $\mathtt{MCB}$ (b) $\mathtt{MCTFE}$ (c) $\mathtt{MCFE}$

(d) $\mathtt{MCB\_l_1}$ (e) $\mathtt{MCTFE\_l_1}$ (f) $\mathtt{MCFE\_l_1}$

Figure 1: MAPE at increasing percentages of unknown entries. Median over the 200 replications (solid lines), 95% confidence bands (blue dashed lines). Top: raw matrix. Bottom: l_1 row-normalization by country.

Counterfactual Analysis



$\mathtt{MCFE}$ on by l_1 country normalized values is applied to estimate the counterfactual CO_2 emissions on the test set (around 50% of total entries).

To draw best and worst case scenario, we represent, for each treated country, **10th, 50th and 90th percentiles** from **80 replications** with randomly selected different training and validation sets.

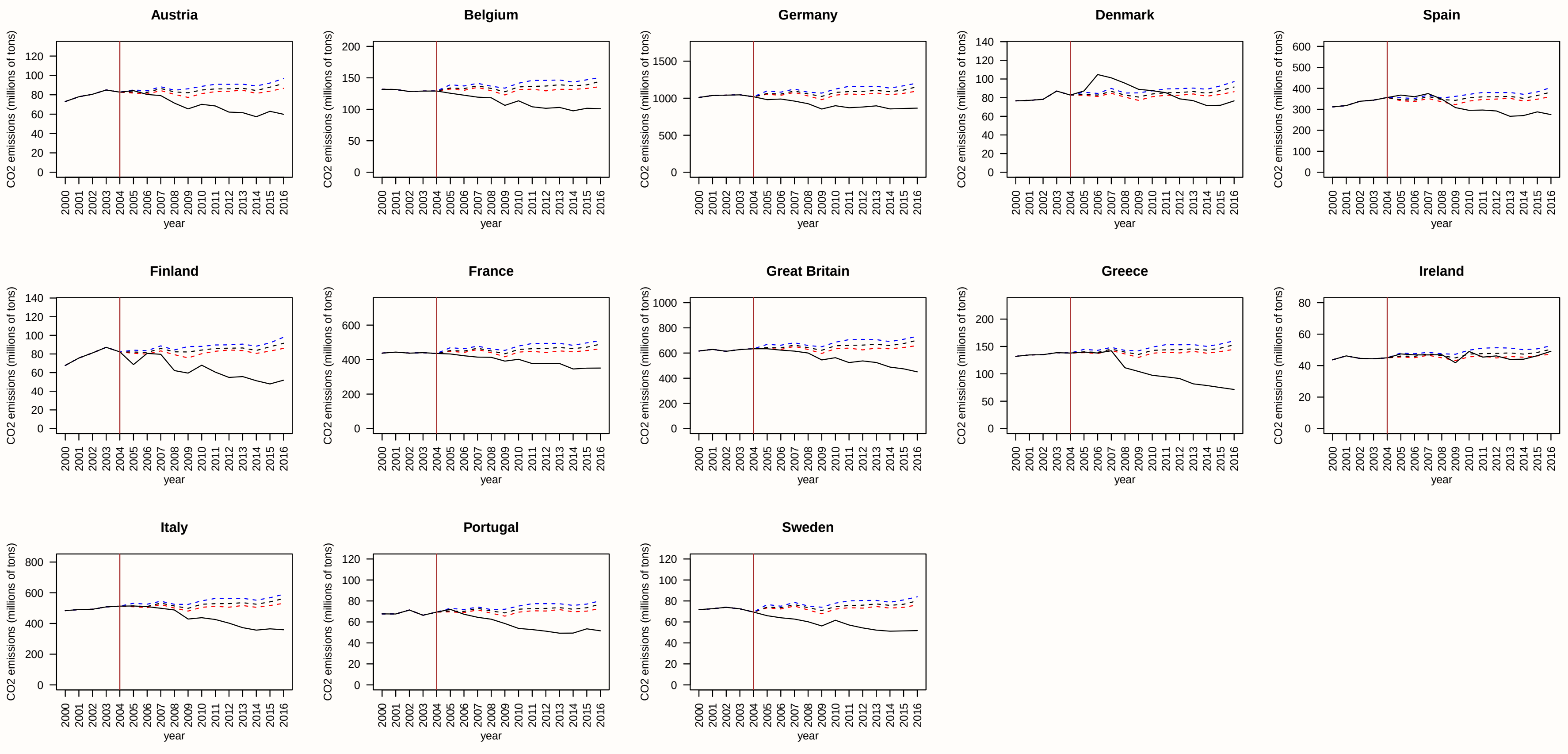


Figure 2: Total CO_2 emissions of treated countries. Actual values (black lines) compared to counterfactual values calculated by MCFE (test set). Medians (black dashed lines), 10th percentiles (red dashed lines), and 90th percentiles (blue dashed lines) considering the 80 MCFE random simulations. Vertical red lines divide the period into pre-treatment and treatment.

SIS 2023 - Statistical Learning, Sustainability and Impact Evaluation, June 21 – 23, 2023, Ancona, Italy