Optimization Letters
https://doi.org/10.1007/s11590-023-02052-2

ORIGINAL PAPER

Prediction of annual CO_2 emissions at the country and sector levels, based on a matrix completion optimization problem

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

Received: 21 August 2022 / Accepted: 27 July 2023

Abstract

In the recent past, annual CO_2 emissions at the international level were examined from various perspectives, motivated by rising concerns about pollution and climate change. Nevertheless, to the best of the authors' knowledge, the problem of dealing with the potential inaccuracy/missingness of such data at the country and economic sector levels has been overlooked. Thereby, in this article we apply a supervised machine learning technique called Matrix Completion (MC) to predict, for each country in the available database, annual CO_2 emissions data at the sector level, based on past data related to all the sectors, and more recent data related to a subset of sectors. The core idea of MC consists in the formulation of a suitable optimization problem, namely the minimization of a proper trade-off between the approximation error over a set of observed elements of a matrix (training set) and a proxy of the rank of the reconstructed matrix, e.g., its nuclear norm. In the article, we apply MC to the imputation of (artificially) missing elements of country-specific matrices whose elements come from annual CO_2 emission levels related to different sectors, after proper pre-processing at the sector level. Results highlight typically a better performance of the combination of MC with suitably-constructed baseline estimates with respect to the baselines alone. Potential applications of our analysis arise in the prediction of currently missing elements of matrices of annual CO_2 emission levels and in the construction of counterfactuals, useful to estimate the effects of policy changes able to influence the annual CO_2 emission levels of specific sectors in selected countries.

Keywords Machine learning · Regularized optimization · Matrix completion · Prediction · Pollution

Extended author information available on the last page of the article

Published online: 26 September 2023

1 Introduction

The concern about pollution and climate change is rising every year with media and public opinion is increasingly concerned about global sustainability. Certainly, one reason of concern is the annual emission of a huge amount of carbon dioxide (CO_2), largely derived from anthropogenic activities like transportation, heavy industries, and electricity generation from fossil combustibles. For these reasons, at the international level, several countries and supranational organizations are devising strategies to decrease the consumption of hydrocarbons, a notable example being the Paris Agreement in 2016, whose main goal is to reduce the annual CO_2 emission levels by at least 55% by 2030 (compared to their values reached in 1990). The relevance of the problem is highlighted by the fact that international agreements on global CO_2 emissions reduction involve countries only on a voluntary basis. In other words, there is currently an absence of commitment at an international level to pollution control. This shows how it is difficult to reach a global, enforceable agreement on the reduction of CO_2 emission levels (although some theoretical models about the effectiveness of possible commitment policies have been developed in the literature: see, e.g., El Ouardighi et al. [8] and El Ouardighi et al. [9]). The problem is also highly relevant from an economic point of view. In this respect, a key statistic describing climate change impacts of CO_2 emissions is the so-called social cost of CO_2 (see, e.g., Kikstra et al. [18]), i.e., the projected cost to society of releasing an additional ton of CO_2.

Nevertheless, to the best of the authors' knowledge, the potential inaccuracy/missingness of CO_2 emissions data in certain countries was overlooked, at least at the country level. Thereby, taking the hint from the past successful applications of machine learning to environmental sciences (see, e.g., Hsieh [16]), in this article, we contribute to the topic by studying CO_2 emissions by applying a Supervised Machine Learning (SML) method to a country-sector level database spanning several years in the recent past. Specifically, we employ a method called Matrix Completion (MC, see Hastie et al. [15]), which was recently popularized, among others, by the 2021 Nobel-prize winner in Economics, Guido W. Imbens [1]. Its core idea consists in the formulation of a suitable optimization problem modeling SML [23], namely the minimization of a proper trade-off between the approximation error over a set of observed elements of a matrix (training set) and a proxy of the rank of the reconstructed matrix, e.g., its nuclear norm (i.e., the summation of all its singular values). A strong advantage of MC with respect to other methods resides in its flexibility, which permits it to be adopted, with appropriate adaptations, in various fields of research. Classical applications of various forms of MC (see, e.g., Candès and Recht [3]) arise, e.g., in collaborative filtering, system identification, and recovery of sensor maps. Two recent successful examples of MC application are represented by the works Metulini et al. [22], where MC is applied for the reconstruction of World Input–Output Database (WIOD) subtables, and Gnecco et al. [14], where MC is used to define a novel index of economic complexity, based on the different degree of predictability of the elements of the Revealed Comparative Advantage (RCA) matrix which are

associated with each country. In the environmental context, an application of MC to climate prediction is made by Ghafarianzadeh and Monteleoni [10]. A recent use of MC for the prediction of CO_2 emission levels is made by Huang et al. [17], but is limited to 11 urban areas in China. Such an application is justified therein by the presence of locally missing data on CO_2 emission levels.

Given the framework above, in this work we propose a specific adjustment (i.e., a proper pre-processing of the available dataset) that may improve MC performance for its specific use with data associated with annual CO_2 emission levels on a country and economic sector basis. More in details, we apply MC to country-specific subsets of the available pre-processed database of CO_2 emissions, combining its prediction with the ones of suitably-constructed baselines. The statistical significance of the results of the one-sided Wilcoxon matched-pairs signed-rank tests performed highlights that the combination of MC with the baselines has typically a better performance than the baselines themselves. The present article represents a thorough extension of the short conference article by the same authors (Biancalani et al. [2]), in which MC was compared only with a simple baseline (the sector-specific average over all the available years in the training set) and no MC/baseline combination was performed.

The article is structured as follows. Section 2 provides a description of the database used for the successive analysis. Section 3 details the adopted methodology of analysis, which is based on a regularized matrix completion optimization problem. Section 4 reports its results. Finally, Sect. 5 concludes the work, pointing out possible extensions of the analysis.

2 Database description

The database used for our analysis is described by Corsatea et al. [4], and is freely downloadable from the following hyperlink: https://joint-research-centre.ec.europa.eu/document/download/b572c87b-a2fb-4ab6-af38-ff0451273e9e_en?filename=co2em56.zip. It refers to annual CO_2 emission levels from 56 economic sectors and from households, for 12 energy commodities. It covers 29 European countries and 13 other major countries in the world, in the period 2000–2016 (one observation for each country, sector, and year). Namely, the 42 countries which are covered in the database are the following ones: AUS (Australia), AUT (Austria), BEL (Belgium), BGR (Bulgaria), BRA (Brazil), CAN (Canada), CHE (Chile), CHN (China), CYP (Cyprus), CZE (Czech Republic), DEU (Germany), DNK (Denmark), ESP (Spain), EST (Estonia), FIN (Finland), FRA (France), GBR (Great Britain), GRC (Greece), HRV (Croatia), HUN (Hungary), IDN (Indonesia), IND (India), IRL (Ireland), ITA (Italy), JPN (Japan), KOR (South Korea), LVA (Latvia), LTU (Lithuania), LUX (Luxembourg), MEX (Mexico), MLT (Malta), NOR (Norway), POL (Poland), PRT (Portugal), ROU (Romania), RUS (Russia), SVK (Slovakia), SVN (Slovenia), SWE (Sweden), TUR (Turkey), TWN (Taiwan), and USA (United States of America).

In our analysis, for each country, data associated with the last three rows of the corresponding matrix of CO_2 yearly emission levels are removed. They correspond, respectively, to the emissions associated with 2 specific sectors (coded, respectively, as *T*: "Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use", and *U*: "Activities of extraterritorial organizations and bodies"), for which the annual CO_2 emission levels reported in the database are typically 0 or nearly equal to 0; to the emissions associated with final consumption expenditure by households (coded as *FC_HH*). Indeed, the interest of the present analysis is in the emissions related to production activities. Concluding, a total of 54 sectors is considered to perform our analysis. The resulting yearly CO_2 emission levels matrices (one for each country considered) have 54 rows and 17 columns.

3 Methodology

In the article, we exploit the following formulation of the Matrix Completion (MC) optimization problem, which was studied theoretically by Mazumder et al. [21]:

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

where Ω^{tr} (which, using a machine-learning expression, can be called training set) is a subset of pairs of indices (i, j) corresponding to positions of known elements of a matrix $\mathbf{M}\in\mathbb{R}^{m\times n}$, $\hat{\mathbf{M}}$ is the completed matrix (to be optimized by solving the optimization problem above), $\lambda\geq 0$ is a regularization parameter, and $\|\hat{\mathbf{M}}\|_*$ is the nuclear norm of the matrix $\hat{\mathbf{M}}$. The problem has a similar structure as the well-known Least Absolute Shrinkage and Selection Operator (LASSO) optimization problem (see, e.g., Kim and Bou [19], for its short presentation). The regularization parameter λ controls the trade-off between fitting the known elements of the matrix $\mathbf{M}$ and achieving a small nuclear norm of its reconstruction $\hat{\mathbf{M}}$. In this article, the optimization problem (1) is solved numerically by applying an iterative algorithm called Soft Impute, developed by Mazumder et al. [21] – to which we refer for a convergence analysis – and reported in the following as Algorithm 1 (see also the supplementary material of Gnecco et al. [14] for a short discussion of implementation issues about this algorithm). The following notation is used. For a matrix $\mathbf{Y}\in\mathbb{R}^{m\times n}$, $\mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{Y})$ represents the projection of $\mathbf{Y}$ onto Ω^{tr}, $\mathbf{P}^{\perp}_{\Omega^{\text{tr}}}(\mathbf{Y})$ denotes the projection of $\mathbf{Y}$ onto the complement of Ω^{tr}, whereas $\mathbf{S}_{\lambda}(\mathbf{Y}):=\mathbf{U}\mathbf{\Sigma}_{\lambda}\mathbf{V}^{\top}$, being $\mathbf{Y}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$ (with $\mathbf{\Sigma}=\text{diag}[\sigma_1,\ldots,\sigma_R]$) the singular value decomposition of $\mathbf{Y}$, and $\mathbf{\Sigma}_{\lambda}:=\text{diag}[(\sigma_1-\lambda)_+,\ldots,(\sigma_R-\lambda)_+]$, with $t_+:=\max(t,0)$.

Algorithm 1: Soft Impute (Mazumder et al., 2010 [21])

Input: Partially observed matrix $\mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{M})$, regularization constant $\lambda \geq 0$, tolerance $tol \geq 0$, maximal number of iterations N^{it}

Output: Completed matrix $\hat{\mathbf{M}} \in \mathbb{R}^{m \times n}$

1. Initialize $\hat{\mathbf{M}}$ as $\hat{\mathbf{M}}^{\text{old}} = \mathbf{0} \in \mathbb{R}^{m \times n}$
2. Repeat for at most N^{it} iterations:
   (a) Set $\hat{\mathbf{M}}^{\text{new}} \leftarrow \mathbf{S}_{\lambda}\left(\mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{M}) + \mathbf{P}^{\perp}_{\Omega^{\text{tr}}}(\hat{\mathbf{M}}^{\text{old}})\right)$
   (b) If $\frac{\|\hat{\mathbf{M}}^{\text{new}} - \hat{\mathbf{M}}^{\text{old}}\|_F^2}{\|\hat{\mathbf{M}}^{\text{old}}\|_F^2} < tol$, exit
   (c) Set $\hat{\mathbf{M}}^{\text{old}} \leftarrow \hat{\mathbf{M}}^{\text{new}}$
3. Set $\hat{\mathbf{M}} \leftarrow \hat{\mathbf{M}}^{\text{new}}$

At each iteration, the Soft Impute algorithm exploits its current solution (called $\hat{\mathbf{M}}^{\text{old}}$) to the MC optimization problem (1) with the aim of estimating the unobserved portion of the matrix $\mathbf{M}$. This estimate is used to generate a completed matrix $\hat{\mathbf{M}}$ which, differently from $\hat{\mathbf{M}}^{\text{old}}$, coincides by construction with $\mathbf{M}$ on the training set. Finally, a new estimate (called $\hat{\mathbf{M}}^{\text{new}}$) is obtained by computing the singular value decomposition of $\hat{\mathbf{M}}$, reducing by λ its singular values larger than λ, and zeroing all the other singular values. The matrix $\hat{\mathbf{M}}^{\text{new}}$ replaces $\hat{\mathbf{M}}^{\text{old}}$ at the next iteration. The algorithm is initialized with $\hat{\mathbf{M}}^{\text{old}} = \mathbf{0}$.

In our application, the tolerance parameter of the Soft Impute algorithm (which refers to the minimum allowable relative change $\frac{\|\hat{\mathbf{M}}^{\text{new}} - \hat{\mathbf{M}}^{\text{old}}\|_F^2}{\|\hat{\mathbf{M}}^{\text{old}}\|_F^2}$ of the square of the Frobenius norm of the reconstructed matrix, and forms the termination criterion used by the algorithm) is selected as $tol = 10^{-10}$. Additionally, when convergence is not achieved, in order to reduce the computational effort, the algorithm terminates after $N^{\text{it}} = 10^5$ iterations. This is motivated by the fact that, in our application, MC has to be performed several times, for different matrices $\mathbf{M}$ (one for each country in the database), several training sets Ω^{tr}, and various choices of λ.

Since MC typically achieves better performance when the elements of the matrix to which it is applied have similar orders of magnitude (see, e.g., its successful application considered by Gnecco et al. [12], where the matrix elements are percentages between 0% and 100%), for every country, the original matrix of annual CO_2 emissions is pre-processed by dividing every row by the l_1 norm of that row restricted to the training set (i.e., by the summation of the absolute values of its elements restricted to the training set), then multiplying it by the fraction of observed elements in that row (this pre-processing step is not performed in case a row contains only zeros in the associated training set). The resulting (country-specific) matrix is denoted as $\mathbf{M}^{\text{pre-processed}}$. It is worth observing that such a pre-processing exploits only the elements of the training set (i.e., the elements of the validation and test sets, which are described later in this section, are

not involved in such a step). In other words, this pre-processing aims at making similar the orders of magnitude of the elements belonging to different rows of the pre-processed matrix, and at the same time it does not use any information that one may want to predict later using MC, since this would be unfair.

Then, we consider the three following methods:

1. In the first case, MC is applied directly to the pre-processed matrix, i.e., one takes $\mathbf{M} = \mathbf{M}^{\text{pre-processed}}$ in Eq. (1). The output $\hat{\mathbf{M}}$ of Algorithm 1 is then taken as an estimate of the pre-processed matrix, i.e., one takes $\hat{\mathbf{M}}^{\text{pre-processed}} = \hat{\mathbf{M}}$. In the following, this method is denoted simply as "MC".
2. In the second case, first a suitable baseline estimate is generated, which is collected in a matrix $\hat{\mathbf{M}}^{\text{baseline}}$. Then, MC is applied to the difference between the pre-processed matrix and the baseline estimate matrix, i.e., one takes $\mathbf{M} = \mathbf{M}^{\text{residual}} := \mathbf{M}^{\text{pre-processed}} - \hat{\mathbf{M}}^{\text{baseline}}$ in Eq. (1). The output $\hat{\mathbf{M}}$ of Algorithm 1 is then taken as an estimate of the residual of the pre-processed matrix, i.e., one takes $\hat{\mathbf{M}}^{\text{residual}} = \hat{\mathbf{M}}$, or equivalently $\hat{\mathbf{M}}^{\text{pre-processed}} = \hat{\mathbf{M}}^{\text{baseline}} + \hat{\mathbf{M}}$. In the following, this method is denoted as "MC/baseline".
3. In the third case, only the baseline estimate is used, hence one gets $\hat{\mathbf{M}}^{\text{pre-processed}} = \hat{\mathbf{M}}^{\text{baseline}}$. In the following, this method is denoted as "baseline".

In both cases 1 and 2, an additional post-processing step is included, thresholding to 0 any negative element (when present) of the matrix $\hat{\mathbf{M}}^{\text{pre-processed}}$. This step is motivated by the fact that the original matrix of annual CO_2 emission levels is non-negative (likewise its pre-processed version). Such a step is not needed for case 3, assuming that the baseline estimates are non-negative (as it occurs for the choices of the baselines detailed in the following). In both cases 1 and 2, for every λ, the resulting completed and thresholded matrix is denoted as $\hat{\mathbf{M}}_{\lambda}^{\text{pre-processed}}$.

In the following, three different baseline estimate matrices are considered, denoted as $\hat{\mathbf{M}}^{\text{baseline}_1}$, $\hat{\mathbf{M}}^{\text{baseline}_2}$, and $\hat{\mathbf{M}}^{\text{baseline}_3}$. They are defined, respectively, as follows:

- each element of $\hat{\mathbf{M}}^{\text{baseline}_1}$ is generated as the sector-specific (i.e., row-specific) simple moving average (Chiulli [5]) of $\mathbf{M}^{\text{pre-processed}}$ over the previous 5 years in the training set (respectively, for the first 5 years, the value itself on each element of the training set);
- each element of $\hat{\mathbf{M}}^{\text{baseline}_2}$ is generated as the year-specific (i.e., column-specific) average of $\mathbf{M}^{\text{pre-processed}}$ over the training set (by construction, $\hat{\mathbf{M}}^{\text{baseline}_2}$ is constant on each column);
- $\hat{\mathbf{M}}^{\text{baseline}_3}$ is the mean of $\hat{\mathbf{M}}^{\text{baseline}_1}$ and $\hat{\mathbf{M}}^{\text{baseline}_2}$, i.e., $\hat{\mathbf{M}}^{\text{baseline}_3} = \frac{\hat{\mathbf{M}}^{\text{baseline}_1} + \hat{\mathbf{M}}^{\text{baseline}_2}}{2}$.

The choice of the first baseline is motivated by the fact that a preliminary descriptive analysis highlighted that, for every country, annual CO_2 emission levels of every sector change typically quite smoothly from one year to the successive one. The choice of the second baseline is motivated by the fact that some yearly changes

across sectors can be still observed, especially between the years 2008 and 2009. However, annual CO_2 emission levels among sectors are still quite heterogeneous at the cross-sectional level. For this reason, the prediction capability of the second baseline is expected to be smaller than that of the first baseline. An intermediate case is represented by the third baseline, which combines the first two baselines.

The reason for which one expects that applying MC on the residual of a baseline estimate matrix improves the performance of MC when the latter is used alone – at least when the baseline has good prediction capability – is that, in the MC optimization problem (1), the non-negative regularization term $\lambda\|\hat{\mathbf{M}}\|_*$ refers to the whole completed matrix $\hat{\mathbf{M}}$. Hence, for large λ, the elements of $\hat{\mathbf{M}}$ tend to be shrunk towards 0, in a similar way as it occurs in the case of the LASSO optimization problem. In particular, this can cause a negative bias in the estimates when $\mathbf{M}$ is a matrix with non-negative elements (moreover, biasedness can be ascribed also to the fact that the Soft Impute algorithm is initialized with a $\mathbf{0}$ matrix, and terminates after N^{it} iterations). The combination of MC with a suitable baseline estimate matrix is expected to make such a bias less negative (and possibly improve the MC performance), because in this case the reconstructed matrix is decomposed into the sum of two matrices, and the regularization term acts not on the whole reconstructed matrix, but only on one of these two matrices. Finally, the combination of MC with a baseline is expected to have better generalization capability than the baseline itself, since in that case, for $\lambda = 0$, the optimal solution to the MC optimization problem (1) does not alter the baseline estimate outside the training set, whereas improvements are likely to obtained for other values of λ.

In the present MC application to (country-specific) matrices associated with CO_2 emissions, the union of the validation and test sets refers to positions of matrix elements which are artificially obscured (but are still available as a ground truth), whereas the training set refers to the positions of all the remaining elements of the matrix considered. Specifically, for every country, MC is applied 20 times, every time generating the training set as follows:

- 50% of randomly selected rows (sectors) are observed over the whole time period;
- the remaining 50% rows are observed over all the years, except for the last three years in the database (namely, 2014, 2015, and 2016).

Then, to avoid overfitting, the regularization parameter λ is chosen according to the following validation method. First, the set of positions of unobserved elements of the matrix $\mathbf{M}$ is partitioned randomly into a validation set Ω^{val} (which contains about 25% of the positions of the unobserved elements) and a test set Ω^{test} (which refers to the positions of the remaining elements). In order to ease the comparison of the MC results when considering different countries, the random choices of the training, validation, and test sets are the same for every country, in each of the 20 repetitions of the MC application (nevertheless, distinct repetitions turn out to have different realizations of the training, validation, and test sets). It is worth noting that, by construction, the training, validation, and test sets do not overlap.

Finally, the optimization problem (1) is solved for several choices λ_k for λ, exponentially distributed as $\lambda_k = 2^{k/2-25}$, for $k = 1, \ldots, 100$. For every λ_k, the Root Mean Square Error (RMSE) of matrix reconstruction on the validation set is computed as

$$RMSE^{\text{val}}_{\lambda_k} := \sqrt{\frac{1}{|\Omega^{\text{val}}|} \sum_{(i,j)\in\Omega^{\text{val}}} \left(\hat{M}^{\text{pre-processed}}_{\lambda_k,i,j} - M^{\text{pre-processed}}_{i,j}\right)^2}, \tag{2}$$

then the choice λ_{k° that minimizes $RMSE^{\text{val}}_{\lambda_k}$ for $k = 1, \ldots, 100$ is obtained. For every λ_k, the RMSEs of matrix reconstruction on the training and test sets ($RMSE^{\text{tr}}_{\lambda_k}$ and $RMSE^{\text{test}}_{\lambda_k}$) are defined similarly, as

$$RMSE^{\text{tr}}_{\lambda_k} := \sqrt{\frac{1}{|\Omega^{\text{tr}}|} \sum_{(i,j)\in\Omega^{\text{tr}}} \left(\hat{M}^{\text{pre-processed}}_{\lambda_k,i,j} - M^{\text{pre-processed}}_{i,j}\right)^2}, \tag{3}$$

and

$$RMSE^{\text{test}}_{\lambda_k} := \sqrt{\frac{1}{|\Omega^{\text{test}}|} \sum_{(i,j)\in\Omega^{\text{test}}} \left(\hat{M}^{\text{pre-processed}}_{\lambda_k,i,j} - M^{\text{pre-processed}}_{i,j}\right)^2}. \tag{4}$$

In the following section, focus is given to their values obtained for $\lambda = \lambda_{k^\circ}$.

To conclude, it is worth discussing at least shortly some computational aspects. Achieving an optimal selection for the λ parameter in the MC optimization problem (1) can be very expensive from a computational point of view, and several methods were proposed in the specialized literature to accelerate the Soft Impute algorithm, e.g., by approximating its singular value thresholding phase, or by introducing a warm-start phase, which initializes λ near its optimal value: see, e.g., Yao and Kwok [24] and de Araújo et al. [6]. Constrained variations of the optimization problem (1) were also proposed (see, e.g., Duarte et al. [7] for some examples). However, for our analysis, the selection of an optimal λ (one selection for each country/repetition pair) is not particularly time-consuming, due to the small size of the matrices considered (indeed, each matrix is made of 54 rows and 17 columns). Nevertheless, for bigger matrices, observing some stability of the optimal λ parameter with respect to changing repetition or changing the country under investigation can help speeding up substantially the application of MC. A final remark has to do with how to make the results of the analysis reproducible, since the choices of the training, validation, and test sets are random. This aim can be easily fulfilled by using deterministic algorithms for sampling (see, e.g., Gnecco et al. [11]). In the specific case, pseudo-random number generators can be employed to generate such sets in a deterministic way. The results reported in the next section refer, indeed, to the case of pseudo-random number generation.

4 Results

For the reasons outlined in Sect. 3, for each choice of the baseline, the following outcomes are expected, when comparing the three methods ("MC/baseline", "MC", "baseline"):

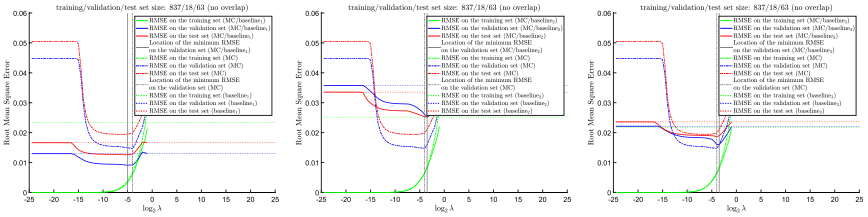
1. "MC/baseline" is expected to have a larger prediction performance than "baseline" alone;
2. "MC/baseline" is expected to have a larger prediction performance than "MC" alone, when "baseline" has good prediction performance;
3. "MC" alone is expected to have a larger prediction performance than "baseline" alone, when "baseline" has little prediction performance.

In the following, results that confirm statistically these expectations are provided.

To begin with the presentation of the outcomes of the analysis, Figs. 1 and 2 illustrate the results obtained in one repetition of the analysis, for a representative country (Spain), taken as case study (similar results are achieved for other repetitions, as detailed in the following). In particular, Fig. 1 shows, for each of the three baselines: the RMSEs on the training, validation and test sets achieved by the combined MC/baseline method described in Sect. 3 (solid colored lines); the RMSEs on the training, validation and test sets achieved by the MC method alone (dash-dotted colored lines); their comparison with the respective RMSEs produced by the baseline alone (dashed horizontal colored lines); the location of the minimum RMSE on the validation set, for the combined MC/baseline method (solid vertical black line) and for the MC method (dash-dotted vertical black line). Moreover, for illustrative purposes, Fig. 2 provides, just for the case of the combined MC/baseline method and the first baseline, a colored visualization of: the elements of the original matrix (to be reconstructed); the positions of its missing elements associated with the specific repetition; the reconstructed matrix obtained in correspondence of the optimal choice of the regularization parameter; the respective element-wise absolute value of the reconstruction error. It is worth observing that, for baseline_1, data related to the first five years are reconstructed exactly by the combined MC/baseline method, since the corresponding columns in the training set used by MC are made exclusively by zeroes, due to the removal of the specific baseline from the matrix to be reconstructed.

Then, Table 1 reports, for all the 20 repetitions and the associated test sets, the RMSE on the test set for the combined MC/baseline method (in correspondence of the optimal choice of the regularization parameter λ) for the representative country and each baseline; the RMSE on the same test set for the MC method (in correspondence of the optimal choice of the regularization parameter λ); the RMSE on the same test set obtained by each baseline.

1. In the case of the representative country, for each baseline, the combined MC/baseline method shows a statistically significant better performance than the baseline alone. Indeed, the application of a one-sided Wilcoxon matched-pairs signed-



(a) Combination/comparison of MC with baseline$_1$ (b) Combination/comparison of MC with baseline$_2$ (c) Combination/comparison of MC with baseline$_3$

Fig. 1 Results of the application of one repetition of the combined MC/baseline method, of MC alone, and of the baseline alone, in the case of a representative country (Spain). **a** Case of baseline$_1$. **b** Case of baseline$_2$. **c** Case of baseline$_3$

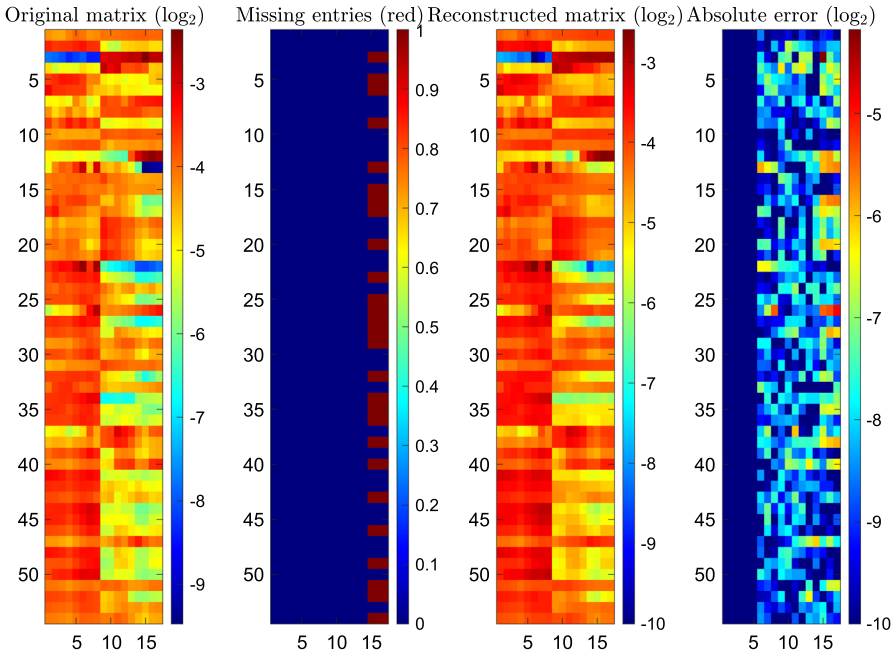


Fig. 2 Colored visualization of: pre-processed elements of the annual CO_2 emission levels matrix related to a representative country (Spain); positions of the missing elements (in both the validation and test sets) in one repetition of the combined MC/baseline$_1$ method; reconstructed matrix obtained in correspondence of the optimal choice of the regularization parameter; respective element-wise absolute value of the reconstruction error. The x-axis refers to the year (the year 2000 being associated with the column number 1, the year 2016 with the column number 17), whereas the y-axis refers to the sector. In the last subfigure, errors smaller than 2^{-10} have been replaced by 2^{-10}, for a better visual representation

rank test (used with an analogous goal as in the work Gnecco and Nutarelli [13]) rejects the null hypothesis that the difference $\Delta^{\text{baseline}}_{\text{MC/baseline}} := RMSE^{\text{test}}(\text{baseline}) - RMSE^{\text{test}}_{\lambda_{k^\circ}}(\text{MC/baseline})$ has a symmetric

Table 1 RMSEs on the test set for the various methods considered in the article, in the case of a representative country (Spain)

	Repetition number									
	1	2	3	4	5	6	7	8	9	10
$RMSE^{test}_{\lambda_{k^\circ}}$ (MC/baseline$_1$)	0.0126	0.0064	0.0088	0.0327	0.0382	0.0073	0.0386	0.0105	0.0371	0.0328
$RMSE^{test}_{\lambda_{k^\circ}}$ (MC/baseline$_2$)	0.0252	0.0165	0.0130	0.0486	0.0531	0.0161	0.0490	0.0241	0.0530	0.0434
$RMSE^{test}_{\lambda_{k^\circ}}$ (MC/baseline$_3$)	0.0190	0.0106	0.0100	0.0397	0.0460	0.0106	0.0443	0.0165	0.0451	0.0381
$RMSE^{test}_{\lambda_{k^\circ}}$ (MC)	0.0199	0.0113	0.0156	0.0501	0.0538	0.0076	0.0492	0.0209	0.0528	0.0443
$RMSE^{test}$ (baseline$_1$)	0.0167	0.0103	0.0125	0.0366	0.0438	0.0111	0.0449	0.0151	0.0438	0.0391
$RMSE^{test}$ (baseline$_2$)	0.0336	0.0234	0.0197	0.0561	0.0628	0.0221	0.0582	0.0344	0.0627	0.0517
$RMSE^{test}$ (baseline$_3$)	0.0236	0.0152	0.0133	0.0447	0.0520	0.0142	0.0506	0.0232	0.0517	0.0444
	Repetition number									
	11	12	13	14	15	16	17	18	19	20
$RMSE^{test}_{\lambda_{k^\circ}}$ (MC/baseline$_1$)	0.0301	0.0097	0.0330	0.0304	0.0116	0.0124	0.0103	0.0083	0.0116	0.0361
$RMSE^{test}_{\lambda_{k^\circ}}$ (MC/baseline$_2$)	0.0393	0.0174	0.0412	0.0391	0.0239	0.0212	0.0160	0.0143	0.0210	0.0497
$RMSE^{test}_{\lambda_{k^\circ}}$ (MC/baseline$_3$)	0.0346	0.0127	0.0367	0.0345	0.0159	0.0150	0.0124	0.0107	0.0155	0.0427
$RMSE^{test}_{\lambda_{k^\circ}}$ (MC)	0.0393	0.0145	0.0397	0.0394	0.0195	0.0180	0.0151	0.0107	0.0198	0.0509
$RMSE^{test}$ (baseline$_1$)	0.0343	0.0140	0.0386	0.0356	0.0155	0.0150	0.0128	0.0118	0.0158	0.0431
$RMSE^{test}$ (baseline$_2$)	0.0478	0.0245	0.0497	0.0471	0.0339	0.0315	0.0242	0.0192	0.0289	0.0583
$RMSE^{test}$ (baseline$_3$)	0.0400	0.0175	0.0431	0.0402	0.0222	0.0206	0.0167	0.0135	0.0206	0.0499

distribution around its median and that this median is smaller than or equal to 0 (for each baseline: p-value $= 4.7846 \cdot 10^{-5}$, significance level $\alpha = 0.05$). More generally, the same null hypothesis is rejected for: 36 among the 42 countries in the case of baseline$_1$; 42 among the 42 countries in the case of baseline$_2$; 41 among the 42 countries in the case of baseline$_3$.

2. In the case of the representative country, the same statistical test as above is used to test the null hypothesis that the difference $\Delta^{\text{MC}}_{\text{MC/baseline}} := RMSE^{\text{test}}_{\lambda_{k^\circ}}(\text{MC}) - RMSE^{\text{test}}_{\lambda_{k^\circ}}(\text{MC/baseline})$ has a symmetric distribution around its median and that this median is smaller than or equal to 0 (the optimal regularization parameter λ_{k° may be different for the two methods). For baseline$_1$, the null hypothesis is rejected (p-value $= 4.7846 \cdot 10^{-5}$, significance level $\alpha = 0.05$). For baseline$_2$, the null hypothesis is not rejected (p-value $=$ 0.9868, significance level $\alpha = 0.05$). For baseline$_3$, the null hypothesis is rejected (p-value $= 1.1787 \cdot 10^{-4}$, significance level $\alpha = 0.05$). More generally, the same null hypothesis is rejected for: 41 among the 42 countries in the case of baseline$_1$; 10 among the 42 countries in the case of baseline$_2$; 39 among the 42 countries in the case of baseline$_3$.
3. In the case of the representative country, the same statistical test as above is used to test the null hypothesis that the difference $\Delta^{\text{baseline}}_{\text{MC}} := RMSE^{\text{test}}(\text{baseline}) - RMSE^{\text{test}}_{\lambda_{k^\circ}}(\text{MC})$ has a symmetric distribution around its median and that this median is smaller than or equal to 0. For baseline$_1$, the null hypothesis is not rejected (p-value $= 0.9997$, significance level $\alpha = 0.05$). For baseline$_2$, the null hypothesis is rejected (p-value $= 4.7846 \cdot 10^{-5}$, significance level $\alpha = 0.05$). For baseline$_3$, the null hypothesis is rejected (p-value $= 0.0191$, significance level $\alpha = 0.05$). More generally, the same null hypothesis is rejected for: 0 among the 42 countries in the case of baseline$_1$; 40 among the 42 countries in the case of baseline$_2$; 18 among the 42 countries in the case of baseline$_3$.

It is also worth remarking that, in the case of the results reported in Fig. 1 for the case of the representative country, the choice of the optimal regularization parameter λ_{k° turns out to depend negligibly on the absence/presence/choice of the baseline combined with MC. Similarly, a small dependence of the choice of λ_{k° is observed with respect to the repetition/the choice of the country analyzed, in the sense that they turn out to be always smaller than 1 (a detailed analysis is not shown, due to space limitations).

Additionally, empirical and standard deviations of the quantities $\Delta^{\text{baseline}}_{\text{MC/baseline}}$, $\Delta^{\text{MC}}_{\text{MC/baseline}}$, and $\Delta^{\text{baseline}}_{\text{MC}}$ are reported in Table 2, for the three baselines and for all the countries considered in the analysis, confirming the findings above. Concluding, the numerical results obtained confirm statistically the expectations reported at the beginning of this section.

It is also worth remarking that, as expected according to Sect. 3, for all the countries the average estimate on the test set increased in almost all the 20 repetitions of the analysis, when moving from the original MC method to each combined MC/baseline method. In other words, there was almost always an increase in the bias (estimated) on the test set, i.e., of the following quantity:

Table 2 Empirical means and standard deviations (over the 20 repetitions) of the quantities $\Delta^{\text{baseline}}_{\text{MC/baseline}} := RMSE^{\text{test}}(\text{baseline}) - RMSE^{\text{test}}_{\lambda_k^o}(\text{MC/baseline})$, $\Delta^{\text{MC}}_{\text{MC/baseline}} := RMSE^{\text{test}}(\text{MC}) - RMSE^{\text{test}}_{\lambda_k^o}(\text{MC/baseline})$, $\Delta^{\text{baseline}}_{\text{MC}} := RMSE^{\text{test}}(\text{baseline}) - RMSE^{\text{test}}_{\lambda_k^o}(\text{MC})$ for the three baselines and for all the countries considered in the analysis

Country	Empirical mean (standard deviation)								
	$\Delta^{\text{baseline}_1}_{\text{MC/baseline}_1}$	$\Delta^{\text{MC}}_{\text{MC/baseline}_1}$	$\Delta^{\text{baseline}_1}_{\text{MC}}$	$\Delta^{\text{baseline}_2}_{\text{MC/baseline}_2}$	$\Delta^{\text{MC}}_{\text{MC/baseline}_2}$	$\Delta^{\text{baseline}_2}_{\text{MC}}$	$\Delta^{\text{baseline}_3}_{\text{MC/baseline}_3}$	$\Delta^{\text{MC}}_{\text{MC/baseline}_3}$	$\Delta^{\text{baseline}_3}_{\text{MC}}$
AUS	0.0010(0.0002)	0.0035(0.0020)	− 0.0025(0.0019)	0.0087(0.0014)	− 0.0003(0.0024)	0.0090(0.0031)	0.0040(0.0006)	0.0022(0.0017)	0.0019(0.0019)
AUT	0.0055(0.0011)	0.0048(0.0004)	0.0007(0.0039)	0.0114(0.0017)	− 0.0020(0.0014)	0.0134(0.0018)	0.0049(0.0009)	0.0026(0.0027)	0.0023(0.0024)
BEL	0.0003(0.0004)	0.0030(0.0029)	− 0.0027(0.0029)	0.0018(0.0006)	0.0013(0.0018)	0.0005(0.0019)	0.0000(0.0002)	0.0028(0.0021)	− 0.0028(0.0021)
BGR	0.0007(0.0008)	0.0046(0.0029)	− 0.0040(0.0030)	0.0063(0.0019)	− 0.0002(0.0024)	0.0065(0.0021)	0.0010(0.0005)	0.0041(0.0030)	− 0.0031(0.0028)
BRA	0.0021(0.0008)	0.0041(0.0027)	− 0.0020(0.0030)	0.0114(0.0020)	− 0.0002(0.0033)	0.0116(0.0039)	0.0042(0.0010)	0.0026(0.0026)	0.0016(0.0027)
CAN	0.0027(0.0003)	0.0050(0.0020)	− 0.0023(0.0020)	0.0067(0.0030)	− 0.0025(0.0050)	0.0092(0.0031)	0.0042(0.0011)	0.0020(0.0026)	0.0023(0.0022)
CHE	0.0003(0.0003)	0.0013(0.0016)	− 0.0010(0.0016)	0.0048(0.0016)	− 0.0001(0.0026)	0.0050(0.0024)	0.0013(0.0005)	0.0014(0.0017)	− 0.0001(0.0017)
CHN	0.0003(0.0013)	0.0079(0.0047)	− 0.0076(0.0038)	0.0073(0.0023)	− 0.0001(0.0035)	0.0074(0.0041)	0.0040(0.0013)	0.0053(0.0031)	− 0.0013(0.0026)
CYP	0.0020(0.0010)	0.0017(0.0032)	0.0003(0.0033)	0.0088(0.0012)	− 0.0019(0.0026)	0.0107(0.0022)	0.0023(0.0008)	0.0014(0.0028)	0.0009(0.0024)
CZE	− 0.0001(0.0001)	0.0047(0.0026)	− 0.0048(0.0026)	0.0070(0.0017)	− 0.0018(0.0027)	0.0088(0.0031)	0.0022(0.0009)	0.0021(0.0016)	0.0001(0.0015)
DEU	− 0.0003(0.0006)	0.0042(0.0023)	− 0.0045(0.024)	0.0084(0.0013)	− 0.0010(0.0023)	0.0093(0.0021)	0.0027(0.0010)	0.0023(0.0020)	0.0004(0.0019)
DNK	0.0021(0.0008)	0.0045(0.0023)	− 0.0024(0.0027)	0.0049(0.0006)	0.0000(0.0017)	0.0050(0.0017)	0.0029(0.0005)	0.0028(0.0020)	0.0001(0.0021)
ESP	0.0046(0.0013)	0.0087(0.0046)	− 0.0041(0.0039)	0.0082(0.0014)	− 0.0016(0.0028)	0.0099(0.0029)	0.0053(0.0012)	0.0041(0.0032)	0.0012(0.0027)
EST	0.0006(0.0007)	0.0124(0.0042)	− 0.0018(0.0045)	0.0066(0.0010)	0.0024(0.0043)	0.0041(0.0044)	0.0029(0.0011)	0.0081(0.0038)	− 0.0052(0.0040)
FIN	0.0012(0.0006)	0.0046(0.0020)	− 0.0034(0.0017)	0.0107(0.0014)	− 0.0016(0.0020)	0.0123(0.0021)	0.0039(0.0008)	0.0024(0.0019)	0.0016(0.0015)
FRA	0.0001(0.0002)	0.0024(0.0015)	− 0.0023(0.0016)	0.0087(0.0012)	− 0.0011(0.0013)	0.0098(0.0016)	0.0027(0.0004)	0.0021(0.0011)	0.0006(0.0011)
GBR	0.0003(0.0002)	0.0029(0.0014)	− 0.0026(0.0014)	0.0026(0.0003)	0.0003(0.0014)	0.0023(0.0013)	0.0013(0.0002)	0.0018(0.0014)	− 0.0005(0.0013)
GRC	0.0028(0.0011)	0.0045(0.0062)	− 0.0017(0.0059)	0.0091(0.0015)	− 0.0020(0.0032)	0.0110(0.0028)	0.0029(0.0016)	0.0024(0.0039)	0.0005(0.0034)
HRV	0.0021(0.0006)	0.0022(0.0018)	− 0.0001(0.0016)	0.0164(0.0014)	− 0.0052(0.0015)	0.0216(0.0019)	0.0072(0.0009)	0.0000(0.0010)	0.0072(0.0010)
HUN	0.0027(0.0005)	0.0055(0.0028)	− 0.0028(0.0027)	0.0048(0.0006)	0.0009(0.0022)	0.0039(0.0024)	0.0034(0.0006)	0.0035(0.0020)	− 0.0001(0.0023)
IDN	− 0.0005(0.0010)	0.0061(0.0060)	− 0.0066(0.0060)	0.0084(0.0015)	0.0005(0.0030)	0.0080(0.0025)	0.0030(0.0013)	0.0051(0.0026)	− 0.0021(0.0024)
IND	0.0016(0.0006)	0.0075(0.0036)	− 0.0059(0.0035)	0.0177(0.0025)	− 0.0009(0.0047)	0.0185(0.0043)	0.0083(0.0013)	0.0050(0.0038)	0.0033(0.0030)
IRL	0.0016(0.0010)	0.0015(0.0094)	− 0.0130(0.0096)	0.0097(0.0015)	− 0.0009(0.0027)	0.0105(0.0029)	0.0051(0.0010)	0.0081(0.0055)	− 0.0030(0.0053)
ITA	0.0011(0.0004)	0.0033(0.0021)	− 0.0022(0.0022)	0.0046(0.0008)	− 0.0013(0.0021)	0.0060(0.0021)	0.0026(0.0005)	0.0016(0.0017)	0.0010(0.0018)

Table 2 (continued)

Country	Empirical mean (standard deviation)								
	$\Delta^{baseline_1}_{MC/baseline_1}$	$\Delta^{MC}_{MC/baseline_1}$	$\Delta^{baseline_1}_{MC}$	$\Delta^{baseline_2}_{MC/baseline_2}$	$\Delta^{MC}_{MC/baseline_2}$	$\Delta^{baseline_2}_{MC}$	$\Delta^{baseline_3}_{MC/baseline_3}$	$\Delta^{MC}_{MC/baseline_3}$	$\Delta^{baseline_3}_{MC}$
JPN	0.0016(0.0009)	0.0018(0.0035)	− 0.0002(0.0031)	0.0064(0.0022)	− 0.0005(0.0025)	0.0069(0.0022)	0.0028(0.0012)	0.0018(0.0025)	0.0010(0.0020)
KOR	0.0004(0.0003)	0.0044(0.0017)	− 0.0040(0.0017)	0.0014(0.0024)	− 0.0014(0.0040)	0.0027(0.0028)	0.0008(0.0007)	0.0024(0.0020)	− 0.0016(0.0019)
LTU	0.0010(0.0012)	0.0079(0.0045)	− 0.0070(0.0043)	0.0065(0.0018)	− 0.0008(0.0035)	0.0074(0.0030)	0.0029(0.0016)	0.0044(0.0033)	− 0.0014(0.0030)
LUX	0.0009(0.0007)	0.0082(0.0076)	− 0.0073(0.0076)	0.0065(0.0022)	− 0.0008(0.0052)	0.0073(0.0028)	0.0023(0.0013)	0.0054(0.0052)	− 0.0032(0.0044)
LVA	0.0007(0.0006)	0.0051(0.0057)	− 0.0045(0.0059)	0.0054(0.0013)	0.0009(0.0020)	0.0045(0.0024)	0.0017(0.0009)	0.0044(0.0039)	− 0.0028(0.0038)
MEX	− 0.0001(0.0003)	0.0007(0.0048)	− 0.0008(0.0046)	0.0027(0.0015)	0.0008(0.0030)	0.0018(0.0024)	0.0005(0.0006)	0.0019(0.0029)	− 0.0014(0.0026)
MLT	0.0027(0.0003)	0.0014(0.0023)	0.0013(0.0021)	0.0023(0.0021)	− 0.0035(0.0044)	0.0058(0.0026)	0.0024(0.0009)	− 0.0002(0.0021)	0.0026(0.0016)
NOR	0.0019(0.0009)	0.0055(0.0047)	− 0.0035(0.0045)	0.0094(0.0017)	− 0.0018(0.0024)	0.0113(0.0017)	0.0053(0.0007)	0.0038(0.0032)	0.0015(0.0029)
POL	0.0027(0.0010)	0.0061(0.0036)	− 0.0035(0.0031)	0.0132(0.0012)	− 0.0038(0.0019)	0.0170(0.0027)	0.0065(0.0008)	0.0030(0.0022)	0.0035(0.0024)
PRT	0.0006(0.0002)	0.0035(0.0019)	− 0.0029(0.0019)	0.0044(0.0004)	0.0003(0.0013)	0.0042(0.0014)	0.0024(0.0003)	0.0026(0.0016)	− 0.0002(0.0015)
ROU	0.0014(0.0008)	0.0062(0.0043)	− 0.0048(0.0041)	0.0173(0.0014)	− 0.0055(0.0024)	0.0228(0.0018)	0.0093(0.0007)	0.0023(0.0026)	0.0070(0.0024)
RUS	0.0003(0.0005)	0.0035(0.0017)	− 0.0032(0.0018)	0.0201(0.0019)	− 0.0002(0.0017)	0.0203(0.0024)	0.0080(0.0009)	0.0023(0.0013)	0.0057(0.0018)
SVK	0.0033(0.0030)	0.0089(0.0065)	− 0.0056(0.0064)	0.0043(0.0014)	0.0007(0.0025)	0.0036(0.0028)	0.0038(0.0002)	0.0056(0.0032)	− 0.0017(0.0044)
SVN	0.0003(0.0011)	0.0037(0.0041)	− 0.0034(0.0043)	0.0066(0.0021)	0.0004(0.0018)	0.0061(0.0023)	0.0006(0.0006)	0.0033(0.0029)	− 0.0028(0.0025)
SWE	0.0022(0.0004)	0.0019(0.0031)	0.0003(0.0029)	0.0119(0.0020)	− 0.0034(0.0022)	0.0153(0.0019)	0.0052(0.0012)	0.0009(0.0031)	0.0043(0.0023)
TUR	0.0004(0.0009)	0.0032(0.0072)	− 0.0028(0.0062)	0.0136(0.0127)	0.0517(0.0109)	− 0.0381(0.0115)	0.0075(0.0078)	0.0282(0.0074)	− 0.0207(0.0090)
TWN	0.0016(0.0007)	0.0067(0.0042)	− 0.0051(0.0041)	0.0056(0.0013)	0.0006(0.0017)	0.0050(0.0025)	0.0034(0.0009)	0.0043(0.0028)	− 0.0010(0.0029)
USA	0.0019(0.0005)	0.0043(0.0021)	− 0.0024(0.0021)	0.0040(0.0005)	0.0004(0.0014)	0.0036(0.0014)	0.0025(0.0004)	0.0027(0.0018)	− 0.0001(0.0016)

$$\text{estimated bias}^{\text{test}}_{\lambda_{k^\circ}} := \frac{1}{|\Omega^{\text{test}}|} \sum_{(i,j)\in\Omega^{\text{test}}} \left(\hat{M}^{\text{pre-processed}}_{\lambda_{k^\circ},i,j} - M^{\text{pre-processed}}_{i,j} \right). \tag{5}$$

As expected, for the original MC method, this estimated bias was almost always negative. Again, additional details are not reported, due to space limitations.

5 Conclusions

In the work, matrix completion has been combined and compared with suitable baselines for the reconstruction of artificially missing elements of matrices representing annual CO_2 emissions at the sector level, for each country considered in the analysis. Nevertheless, for specific countries (different from the ones in the database: e.g., selected developing countries), the corresponding matrices may have really missing elements, which could be effectively reconstructed by matrix completion. In this application, the only difference with respect to the current analysis is that it would be not possible to evaluate the error on every element of the test set (being a ground truth unavailable for really missing elements).

A second potential extension of our analysis is represented by the construction of counterfactuals (e.g., as investigated by Kumar and Liang [20], which refers to a different application of matrix completion), useful to predict policy effects on annual CO_2 emissions of specific sectors in selected countries. In practice, this would entail obscuring matrix entries affected by a policy (e.g., any national countermeasure aimed to reduce pollution levels related to the economic activity of specific sectors), to predict their corresponding values in the absence of the policy (counterfactual values). This application would require avoiding getting negatively biased estimates of the unobserved matrix elements. Indeed, overestimates may be actually needed to construct valid counterfactuals.

The analysis made in the present work could be also extended as follows. As an additional step, matrix completion could be applied to matrices obtained by combining the information coming from different countries, or by merging the information available on annual trade flows and annual CO_2 emission. Indeed, a significant feature of the database used for our analysis is that its adopted classification of sectors matches the one of the World Input–Output Database (WIOD) tables (https://www.rug.nl/ggdc/valuechain/wiod/wiod-2016-release), whose elements represent annual trade flows from input country-specific sectors to output country-specific industries/final consumption sectors. Finally, other baselines could be also considered for the combination/comparison with matrix completion.

Acknowledgements The authors acknowledge partial support from the "Dipartimento di Eccellenza 2023–2027" project at IMT - School for Advanced Studies, Lucca. F. Biancalani, G. Gnecco, and R. Metulini are members of GNAMPA (Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni) at INdAM (Istituto Nazionale di Alta Matematica).

Funding Open access funding provided by Scuola IMT Alti Studi Lucca within the CRUI-CARE Agreement.

Data availability The datasets analyzed in the current study are available for free at https://joint-research-centre.ec.europa.eu/document/download/b572c87b-a2fb-4ab6-af38-ff0451273e9e_en?filename=co2em56.zip. This hyperlink was accessed in 2023. The original 2019 version of the database, available at https://joint-research-centre.ec.europa.eu/system/files/2019-09/co2em56.zip, contained also data related to the Netherlands (NLD), which were removed in the updated version of the database.

References

1. Athey, S., Imbens, G.W.: Machine learning methods that economists should know about. Ann. Rev. Econ. **11**, 685–725 (2019)
2. Biancalani, F., Gnecco, G., Metulini, R., Riccaboni, M.: Matrix Completion for the Prediction of Yearly Country and Industry-Level CO_2 Emissions. In Proceedings of the 8th International Conference on machine Learning, Optimization & Data science (LOD 2022), Lecture Notes in Computer Science, vol. 13810, pp. 14-19, (2023)
3. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**, 717–772 (2009)
4. Corsatea, T.D., Lindner, S., Arto, I., Román, M.V., Rueda-Cantuche, J.M., Velázquez Afonso, A., Amores, A.F., Neuwahl, F.: World Input-Output Database Environmental Accounts. Update 2000-2016. Publications Office of the European Union, Luxembourg, (2019) https://doi.org/10.2791/947252
5. Chiulli, R.M.: Quantitative Analysis: an Introduction. CRC Press, (2020)
6. de Araújo, T., Gonçalves, D.S., Torezzan, C.: A two-phase rank-based algorithm for low-rank matrix completion. Opt. Lett. (2022). https://doi.org/10.2791/947252
7. Duarte, L.T., Mussio, A.P., Torezzan, C.: Dealing with missing information in data envelopment analysis by means of low-rank matrix completion. Ann. Oper. Res. **286**, 719–732 (2020)
8. El Ouardighi, F., Kogan, K., Gnecco, G., Sanguineti, M.: Commitment-based equilibrium environmental strategies and time-dependent absorption efficiency. Group Decis. Negot. **27**, 235–249 (2018)
9. El Ouardighi, F., Kogan, K., Gnecco, G., Sanguineti, M.: Transboundary pollution control and environmental absorption efficiency management. Ann. Oper. Res. **287**, 653–681 (2020)
10. Ghafarianzadeh, M., Monteleoni, C.: Climate Prediction Via Matrix Completion. In Proceedings of the 17th AAAI Conference on Late-Breaking Developments in the Field of Artificial Intelligence (AAAIWS'13-17), pp. 35-37, (2013)
11. Gnecco, G., Sanguineti, M., Gaggero, M.: Suboptimal solutions to team optimization problems with stochastic information structure. SIAM J. Optim. **22**, 212–243 (2012)
12. Gnecco, G., Landi, S., Riccaboni, M.: Can machines learn creativity needs? an approach based on matrix completion. Ital. Econ. J. (2022). https://doi.org/10.1007/s40797-022-00200-8
13. Gnecco, G., Nutarelli, F.: On the trade-off between number of examples and precision of supervision in machine learning problems. Opt. Lett. **15**, 1711–1733 (2019)
14. Gnecco, G., Nutarelli, F., Riccaboni, M.: A machine learning approach to economic complexity based on matrix completion. Sci. Rep. (2022). https://doi.org/10.1038/s41598-022-13206-0

15. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: the Lasso and its Generalizations. CRC Press, (2015)
16. Hsieh, W.W.: Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels. Cambridge University Press, (2009)
17. Huang, W., Wei, D., Wang, C., Lin, C.: Matrix completion-based prediction analysis in carbon emissions. Int. J. Embedded Syst. **14**, 143–148 (2021)
18. Kikstra, J.S., Waidelich, P., Rising, J., Yumashev, D., Hope, C., Brierley, C.M.: The social cost of carbon dioxide under climate-economy feedbacks and temperature variability. Environ. Res. Lett. **16**(9), 094037 (2021)
19. Kim, S.H., Boukouvala, F.: Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques. Optim. Lett. **14**, 989–1010 (2020)
20. Kumar, A., Liang, C.-H.: Credit constraints and GDP growth: evidence from a natural experiment. Econ. Lett. **181**, 190–194 (2019)
21. Mazumder, R., Hastie, T., Tibshirani, R.: Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. **11**, 2287–2322 (2010)
22. Metulini, R., Gnecco, G., Biancalani, F., Riccaboni, M.: Hierarchical clustering and matrix completion for the reconstruction of world input-output tables. AStA Ad. Stat. Anal. (2022). https://doi.org/10.1007/s10182-022-00448-6
23. Sra, S., Nowozin, S., Wright, S.J.: (Editors), Optimization for Machine Learning. MIT Press, (2012)
24. Yao, Q., Kwok, J.T.: Accelerated Inexact Soft-Impute for Fast Large-Scale Matrix Completion. In Proceedings of the 24th Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), pp. 4002-4008, (2015)

Authors and Affiliations

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

✉ Giorgio Gnecco
giorgio.gnecco@imtlucca.it

Francesco Biancalani
francesco.biancalani@imtlucca.it

Rodolfo Metulini
rodolfo.metulini@unibg.it

Massimo Riccaboni
massimo.riccaboni@imtlucca.it

[1] IMT - School for Advanced Studies, Piazza San Francesco, 19, 55100 Lucca, Italy

[2] University of Bergamo, Via Caniana 2, 24127 Bergamo, Italy