ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Gap filling in air temperature series by matrix completion methods

Benoît Loucheur[1], P.-A. Absil[1] and Michel Journée[2]

1- ICTEAM Institute, UCLouvain
1348 Louvain-la-Neuve - Belgium

2- Royal Meteorological Institute of Belgium
1180 Uccle - Belgium

Abstract. Quality control of meteorological data is an important part of atmospheric analysis and prediction, as missing or erroneous values can degrade the quality of weather and climate information derived from these data. In practice, the presence of missing data in the weather series is quite common and problematic for many uses. We compare the performance of matrix completion methods with the state of the art to solve this missing data problem. The experimental results are carried out using the daily minimum and maximum temperature measurements of the network of weather stations operated by the Royal Meteorological Institute (RMI) of Belgium.

1 Introduction

In Belgium, the Royal Meteorological Institute (RMI) is the national meteorological service that provides weather and climate services based on observations and scientific research. The RMI collects and archives meteorological observations in Belgium since the 19th century. Currently, air temperature is monitored in Belgium in about 30 synoptic automatic weather stations (AWS) as well as in 110 manual climatological stations. In the latter stations, a volunteer observer records every morning at 8 o'clock the daily extreme air temperatures. These observers communicate their measurements either via a paper bulletin or via the Internet. Missing data are quite common for manual stations, e.g., the daily observations may some days not be done or not properly transmitted to the RMI. AWS data can also sometimes be lacking due to technical problems with the sensor or with the acquisition and communication systems.

Missing data in weather series can be an issue for many uses and in particular for the estimation of climate statistics such as climate normals. Climate normals are 30-year averages of a climate variable that are updated every 10 years by all meteorological institutes according to the recommendations of the World Meteorological Organization [1]. For the recent update in 2021 of the Belgian climate normals, the completion of the weather series was an essential preparatory step as few series had complete data from 1 January 1991 to 31 December 2020 (i.e., some weather stations started after 1991 while others ended before 2020, other series combine data from several neighboring stations with possible interruptions in between and, finally, missing observations are possible for all stations in the operation phase).

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

In this work, we compare matrix completion methods with the state of the art to solve the problem of missing data completion in meteorological time series. The aim of the matrix completion methods is to exploit inherent linear relations within the data in order to recover low rank matrices from a limited number of observations. These methods became very popular in 2006 with the competition launched by the Netflix company: *the Netflix Prize*.

2 Matrix Completion

Given a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, the aim of matrix completion is to recover all its entries from a partially observed fraction of them. The set of observed entries of $\mathbf{M}$ is denoted by $\Omega = \{(i,j) : \mathbf{M}_{ij} \text{ is observed}\}$. As defined by Candès and Recht [2], the projection $\mathcal{P}_\Omega$ with respect to a set of matrix indices Ω is the function $\mathcal{P}_\Omega : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ defined by:

$$[\mathcal{P}_\Omega(\mathbf{M})]_{ij} = \begin{cases} \mathbf{M}_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

In [2], it was shown that the recovery of missing values from a low rank matrix $\mathbf{X}$ is possible by solving the rank minimization problem:

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \quad & \operatorname{rank}(\mathbf{X}), \\ \text{s.t.} \quad & \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{M}). \end{aligned} \tag{1}$$

However, this optimization problem is NP-hard to solve. This formulation tries to minimize the ℓ_0 norm of the singular values (i.e., the rank). Fortunately, it is possible to relax the problem by minimizing instead the ℓ_1 norm of the singular values (i.e., the nuclear norm: $\|\cdot\|_*$):

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \quad & \|\mathbf{X}\|_*, \\ \text{s.t.} \quad & \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{M}). \end{aligned} \tag{2}$$

It has also been shown in [2] that, in the context of *exact* low-rank matrix completion, this formulation recovers the original underlying matrix under some assumptions.

A different way to approach the low rank matrix completion problem is to assume that the rank r of the target matrix is known in advance. In this case, the problem can be stated as:

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \quad & \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{X})\|_F^2, \\ \text{s.t.} \quad & \operatorname{rank}(\mathbf{X}) \leq r. \end{aligned} \tag{3}$$

This new formulation seeks a matrix $\mathbf{X}$ of rank at most r that best fits the given data in Ω.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

A common approach to enforce the rank constraint in (3) is to write $\mathbf{X}$ as $\mathbf{UW}$ with $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{W} \in \mathbb{R}^{r\times n}$ yielding:

$$\min_{\substack{\mathbf{U}\in\mathbb{R}^{m\times r}\\ \mathbf{W}\in\mathbb{R}^{r\times n}}} \quad \frac{1}{2}\left\|\mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{UW})\right\|_F^2 + \underbrace{\frac{\lambda_u}{2}\left\|\mathbf{U}\right\|_F^2 + \frac{\lambda_w}{2}\left\|\mathbf{W}\right\|_F^2}_{\text{Regularization terms}}, \tag{4}$$

where the two regularization terms, added to help avoid overfitting, are related to the nuclear norm (see [3]).

In Section 3, we report on the weather data completion results obtained with two matrix completion methods: SoftImpute [4] and RTRMC [5].
SoftImpute is based on the standard problem (2) to which the constraint is slightly modified to take into account the notion of measurement noise:

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \quad & \left\|\mathbf{X}\right\|_*, \\ \text{s.t.} \quad & \mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{X}) \leq \delta. \end{aligned} \tag{5}$$

The problem is rewritten using the Lagrange form to obtain the SoftImpute formulation:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \quad \frac{1}{2}\left\|\mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{X})\right\|_F^2 + \lambda\left\|\mathbf{X}\right\|_*, \tag{6}$$

where $\lambda > 0$ is a regularization parameter.

RTRMC [5] reformulates (4) with a different regularization term as a nested optimization problem $\min_{\mathbf{U}} \min_{\mathbf{W}}$ and exploits the fact that the inner problem depends on $\mathbf{U}$ only through its column space (which belongs to the Grassmann manifold $\mathcal{G}^{m\times r}$). This yields the formulation:

$$\min_{\mathcal{U}\in\mathcal{G}^{m\times r}} \min_{\mathbf{W}\in\mathbb{R}^{r\times n}} \frac{1}{2}\left\|\mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{UW})\right\|_F^2 + \frac{\lambda^2}{2}\left\|\mathcal{P}_{\bar{\Omega}}(\mathbf{UW})\right\|_F^2, \tag{7}$$

where $\bar{\Omega}$ is the complement of the set Ω.

3 Experimental Results

In this section, the considered matrix completion methods are evaluated on meteorological time series with synthetic gaps and compared against two state-of-the-art methods: Inverse Distance Weighting (IDW) and an approach based on Principal Component Analysis (PCA). IDW [6] is a simple and largely used approach that estimates missing values as weighted averages of neighboring station values with weights that decrease with the distance. PCA is a data reduction method used to find temporal and spatial patterns. Inspired by linear regression, PCA can also be used as a method to complete data [7].

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

3.1 Data

The data completion experiments were performed on a dataset provided by the RMI with daily minimum and maximum temperature measurements from 97 weather stations in Belgium for a period of 15 years (2005 to 2019) without any gaps. This dataset was selected as a compromise between the number of available stations and the length of the series (i.e., a longer period could be considered but for fewer stations and, inversely, a larger number of stations with complete time series could be available but for a shorter period).

3.2 Hyperparameter tuning

All the four considered methods present at least one hyperparameter to be tuned to its best value. As shown in Table 1, each hyperparameter has been assigned a set of possible values. Then, we generate all possible combinations of hyperparameters for each method, as in the GridSearchCV model from Scikit-learn [8].

Methods	Hp	Values tested	T_{min}	T_{max}
PCA [7]	N	$\{1, 2, \cdots, 20\}$	6	5
IDW [6]	p	$\{1, 2, \cdots, 6\}$	6	3
SoftImpute [4]	r	$\{1, 2, \cdots, 20\}$	15	15
RTRMC [5]	r	$\{1, 2, \cdots, 20\}$	13	11
	λ	$\{1, 0.1, 0.01, 0.001\}$	0.001	0.001

Table 1: Set of hyperparameters (Hp) for each method with their optimal values for the daily minimum (T_{min}) and maximum (T_{max}) temperature dataset. N is the number of principal components, p is the power value applied to the weight, r is the rank of the model and λ denotes the regularization parameter.

The best set of hyperparameters was determined for each method by Monte Carlo Cross-Validation [9]. The first 10 years of the dataset (i.e., the data from 2005 to 2014) is separated into training and validation sets to tune the hyperparameters, while the rest of the dataset (i.e., the data from 2015 to 2019) is used as test set to evaluate the performance of the considered methods in their optimal configurations. Synthetic gaps are generated randomly in the validation and test sets to replicate as far as possible the characteristics (i.e., frequency, duration) of the gaps actually present in reality when considering all Belgian temperature series over a longer period (e.g., from 1991 to 2020). Following the examples provided in the introduction, gaps of various sizes (i.e., from one month to 3 years) are generated at the beginning, in the middle and at the end of the validation and test sets. The optimal set of hyperparameters for each method is determined as the one that results into the lowest RMSE when averaged over 10 different random splits of the 2005-2014 data into training and validation sets.

For the final evaluation on the test set, 5 random splits are generated ensuring that each weather station has more or less the same number of synthetic missing data overall.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

3.3 Evaluation on the test set

Table 2 shows the values of different measures of accuracy of the algorithms evaluated on the test set. It is important to remember that the main criterion for choosing the best hyperparameter set is the RMSE. The other criteria provide additional useful information on the distribution of the error.

Daily minimum temperature				
	RMSE	MAE	P_5	P_{95}
PCA	1.214	0.899	-2.005	1.905
IDW	1.138	0.821	-1.45	2.137
SoftImpute	**0.8**	0.587	-1.269	1.3
RTRMC	0.932	0.629	-1.301	1.383
Daily maximum temperature				
	RMSE	MAE	P_5	P_{95}
PCA	1.14	0.842	-1.959	1.725
IDW	0.938	0.71	-1.503	1.575
SoftImpute	0.563	0.414	-0.863	0.913
RTRMC	**0.554**	0.409	-0.886	0.861

Table 2: Average scores (in °C) on the test set containing the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), the 5^{th} and 95^{th} percentile of the error distribution (P_5 and P_{95}).

Matrix completion methods give noticeably better results than state-of-the-art methods. Moreover, the performance gap is more noticeable for daily maximum temperatures than for daily minimum temperatures.

For all methods, the completion performance is worse for the daily minimum temperature data than for the daily maximum temperature data. Daily minimum temperature data exhibits indeed less spatial correlation, i.e., the minimum temperature can vary rapidly with the distance in certain meteorological and topographic conditions. This makes this weather variable less predictable and its completion more challenging.

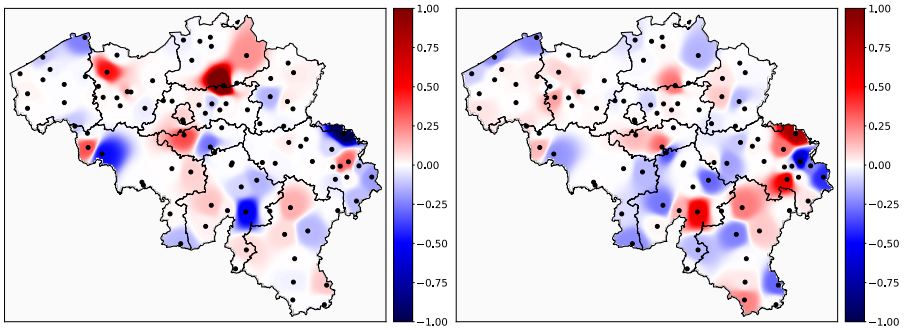
While Table 2 summarizes the performance of the considered methods on average for all stations, the average error from 2015 to 2019 computed separately for each station can be represented on a map as illustrated in Figure 1 for the daily maximum temperature dataset. It is important to note that the scales are not the same for Figure 1a,1b and Figure 1c,1d for the sake of readability.

References

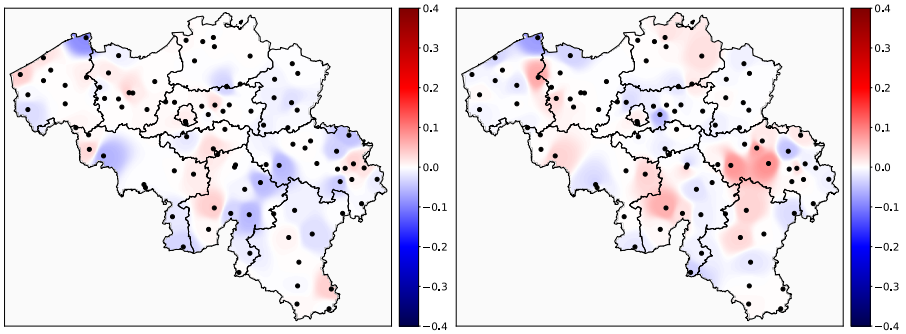
[1] World Meteorological Organization. WMO guidelines on the calculation of climate normals, 2017.

[2] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *CoRR*, abs/0805.4471, 2008.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.



(a) PCA (b) IDW



(c) SoftImpute (d) RTRMC

Fig. 1: Average completion error for each station on the test set (and spatially interpolated by kriging for readability). Case of the daily maximum temperature data. The black dots indicate the position of the stations.

[3] Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research*, 16(104):3367–3402, 2015.

[4] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[5] N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 406–414. 2011.

[6] Arthur T. DeGaetano and Brian N. Belcher. Spatial interpolation of daily maximum and minimum air temperature based on meteorological model analyses and independent observations. *Journal of Applied Meteorology and Climatology*, 46(11):1981 – 1992, 2007.

[7] Hege Hisdal and Ole Tveito. Extension of runoff series using empirical orthogonal functions. *Hydrological Sciences Journal*, 38:33–49, 02 1993.

[8] Steven M LaValle, Michael S Branicky, and Stephen R Lindemann. On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, 23(7-8):673–692, 2004.

[9] Richard R. Picard and R. Dennis Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.