Recovering Missing Photovoltaic Signal (Gaps) With Low Dimensional Manifold Model

Jugurta R. Montalvão Filho[1,2], Leoraig S. Cardoso[1,2], Ezio C. de Santana[1,2], Pierre Pinson[3] and Douglas B. Riffel[1,2]

[1] Federal University of Sergipe, São Cristóvão (Brazil)

[2] INESC P&D Brasil, Santos (Brazil)

[3] Imperial College London (United Kingdom)

Abstract

An adaptation of the Low Dimensional Manifold Model--frequently used for image inpainting--is presented as an alternative for recovering signal (missing data) in photovoltaic generation measurements. This adaptation takes advantage of expected redundancies across simultaneous measurements to fill the gaps by assuming that these redundancies induce low local dimensionality in the mathematical structure (manifold) where multivariate observations are found. Experimental results with actual data from smart meters in Cape Verde corroborate the expected superiority of this approach, as compared to state of the art methods.

Keywords: Signal gaps, LDMM, Inpainting, Reverse Power

1. Introduction

The exponential growth of decentralized photovoltaic (PV) power generation has driven an increase in the number of residential and commercial consumers who have rooftop PV systems using behind-the-meter (BTM) and Net Energy Metering (NEM). This distributed PV generation model has brought new technical and operational challenges, such as gaps in the data gathered from decentralized PV meters, thus making regeneration a crucial step to ensure data quality and enable applications like load/generation forecasting, energy planning, and photovoltaic system performance monitoring.

In that context, Iwueze et al. (2018) compare statistical methods, ranging from simple techniques such as moving averages to more complex approaches like regression and spectral analysis to estimate missing values in time series, whereas Denhard et al. (2021) explore various time-series gap-filling methods specifically for solar irradiance applications. Likewise, Lee and Son (2024) also compare imputation techniques to handle missing PV data, assessing their impact on the performance of PV forecasting models.

In this work, we tackle the problem of data imputation in multivariate NEM, in which local energy consumption plays the role of a stochastic disturbance subtracted from the locally generated power. Our approach is based on the assumption that multivariate NEM signals are redundant, because insolation and circadian energy consumption habits may be shared by neighboring PV installations. This assumption suggests that the Low-Dimensional Manifold Model (LDMM) can be a promising approach, for it exploits redundancies in PV datasets, which induce local low-dimensionality of underlying signal spaces occupied by highly dependent observations, thus eliminating the necessity of non-PV exogenous variables, in contrast with other data imputation methods (Lee and Son, 2024). In other words, the underlying geometric structure of the data is assumed to be locally flat, which favors the use of LDMM, a powerful iterative approach to estimate missing values, as explained in Section 2.

10.18086/swc.2025.04.38 Available at http://proceedings.ises.org

2. Low-Dimensional Manifold Model (LDMM)

The LDMM assumes that high-dimensional redundant data lie on lower-dimensional manifolds within the ambient space. Thus, LDMM exploits this property by iteratively adjusting near sets of observations to conform to a local low-dimensional hyperplane (patches), whereas boundary continuity between patches is iteratively adjusted. Besides, iterations in the LDMM also ensures that the missing parts of the data are reconstructed in a manner that aligns with the learned manifold structure, leading to a sophisticated filtering of data deviations that could not be explained by observed data redundancies.

Montalvão et al. (2024) proposes a straightforward implementation of the LDMM that is used in this work, for it allows an easy adaptation of the LDMM to PV multivariate signals, as follows: let N multivariate observations, from M meters, be represented by a set of real-valued row vectors, corresponding to rows of a $N \times M$ matrix X. Therefore, $X_{n,m}$ stands for the power measurement from meter m at discrete time n.

For a given meter, i.e., a given column of X, gaps (or missing data) typically appear in long bursts, due to communication failures, for instance. For this reason, we adapted the LDMM method to recover data in a row-wise manner, referred to as "horizontal inpainting". In contrast with the more usual "vertical inpainting," in which data from each source (i.e. each column of X) is processed independently, the "horizontal inpainting" takes advantage of redundancies across meters, exploiting the existence of instantaneous cross-dependencies among meters.

Our adaptation of the LDMM performs a total of 30 iterations over each column in X, with parameters experimentally tuned to $\Phi = 0.2$ and $K = 3$. In each iteration, a column of X is scanned and rows of X with gaps in that column are gathered in a temporary matrix G, whereas the remaining rows are gathered in another temporary matrix B. Then, for each row r of G, K nearest columns in B are found through the calculation of the quadratic error (eq. 1) and gathered in yet another matrix C.

$$d_r = \sum_{n=0}^{n=N} (G_r - B_n)^2 \qquad \text{(eq. 1)}$$

where d_r is a vector containing the quadratic error for each column, G_r represents row r of matrix G, and B_n represents row n of matrix B. The K nearest columns are the ones with the smaller values in vector d_r.

Then the calculations in (eq. 2) are performed:

$$W_0 = (CDC^T + I_K)^{-1} CDr^T \qquad \text{(eq. 2)}$$

where I_K stands for the identity matrix and D is a diagonal matrix with entries $|f_m - \Phi|$, $m = 1,2,\dots,M.$, with data flags $f_m \in \{0,1\}$: 1 for available measurement and 0 (gap), otherwise. Thus, with the optimized weights W_0, the m-th missing value in row r is adjusted as in (eq. 3):

$$r_m^{new} \leftarrow \rho r_m^{old} + (1-\rho) c_m^T W_0 \qquad \text{(eq. 3)}$$

where c_m is the m-th column of C.

After all missing values are iteratively regenerated, an additional filter is applied to prevent (spurious) negative values, which may occur as a purely numerical solution to the optimization problem solved by the LDMM.

3. Methodology

The dataset used in this study consisted of Net Energy Metering (NEM) signals from 37 smart meters located in Santiago Island, Cabo Verde. All 37 time series in the dataset spanned a two-week period, with updates every 15 minutes.

To test the inpainting method gaps were synthetically created in each time series through two methods, namely, a completely at random deletion of a selected percentage of real measurements and by using a Markovian memory process. Moreover, for the Markovian memory deletion process, the transition probability matrix chosen for the simulation has entries that stochastically control the proportion of gaps around 50%. More specifically, the 2x2 transition (stochastic) matrix $P = [p_{ij}]$, $i, j \in \{1,2\}$, with $p_{11} = p_{22} = 0.8$ and $p_{12} = p_{21} = 0.2$ is used to simulate instances of the process with equal initial probabilities for both states 1 and 2.

Thus, given an instance of N multivariate observations, from M meters, M instances of the Markovian process, with N observed states, each are simulated and associated with the M meters. To simulate burst of gaps, measurement observations associated to state 2 are replaced with gaps, thus yielding, on average, 50% of gaps independently occurring in each meter, mostly in long bursts, leading to missing data patterns similar to the ones observed in real life, which sometimes happen due to data transmission failures.

The evaluation of the method was done through the calculation of the average coefficient of determination (R^2) between the original data and the data generated for each of the 37 time series. Lastly, the aforementioned testing procedure was repeated 20 times, and the average of the 20 tests was taken as the final result.

$$R^2 = 1 - (\sum_{t=1}^{n}(y_t - y_t')^2) / \left(\sum_{t=1}^{n}(y_t - \overline{y}_t)^2\right) \quad \text{(eq. 4)}$$

where y_t is the real data and y_t' is the generated data.

Furthermore, other than the proposed LDMM method, two other imputation approaches were implemented using the idea of "horizontal inpainting", namely, linear imputation and K-near neighbors. In the linear imputation method, each missing value is calculated as the weighted average of the values from the other meters at that instant, according to (eq. 5):

$$y_t' = \left(\sum_{m=1}^{m=M} y_{t,m}\, e^{-d_{r,m}}\right) / (\sum_{m=1}^{m=M} e^{-d_{r,m}}) \quad \text{(eq. 5)}$$

where $d_{r,m}$ is the quadratic error calculated using (eq. 1).

Lastly, for the KNN method, each missing value is the average of the measurements from the 3 nearest columns, which are again determined using (eq. 1).

4. Results

Table 1 gathers performances of competing methods in terms of R^2, averaged from 20 independent runs, with standard deviations of about 0.02 in all cases.

Table 1: Comparative performances in terms of R2

	R^2				
	Missing Data % (Memoryless)				Missing Data % (Markovian)
Method	**10**	**20**	**30**	**50**	**50**
LDMM	0.96	0.92	0.86	0.75	0.75
KNN	0.95	0.88	0.82	0.67	0.67
Linear	0.95	0.88	0.81	0.65	0.65

Our results are consistent with those found in Lee & Son (2024), in which the KNN method outperforms the linear one. Moreover, in our experiments, the LDMM based method also consistently outperforms the best method used in Lee & Son (2024). A visual comparison among recovered signals from the three methods is shown in Figure 1, for an actual measurement signal with simulated bursts of gaps that occupy 50% of the total durations of observations. As a result of this severe regime of signal degradation, the reconstruction error of unpainted (recovered) signal with respect to the original data is large, as expected, but significantly smaller than the reconstruction error observed from competing methods.

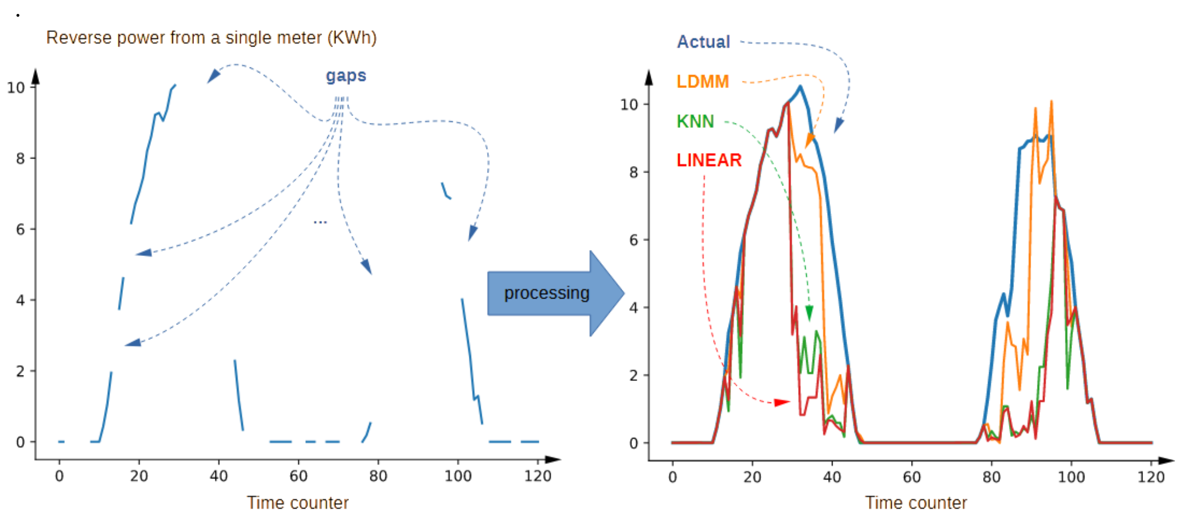

Figure 1: Visual comparison among recovered signals from three methods. The original signal segment (2 day reverse power from a single meter) has not actual gaps, which are simulated as a Markovian process for the purpose of testing method performances. Gaps occupy 50% of the total signal duration.

5. Acknowledgments

The authors gratefully acknowledge the financial support provided by Digital Energy, with funding from the European Union (AfD). This support enabled the successful execution of the DEEP Project in Cape Verde, conducted through a collaborative partnership between INESC P&D Brasil and Electra.

6. Conclusions

A new approach for imputation of missing photovoltaic net power measurements was adapted from the Low-Dimensional Manifold Model. The measurements, combining PV generation and consumption, present a more complex gap- filling challenge due to the combination of two sources of randomness, namely: cloud dynamics and local power consumption. It also introduced a horizontal inpainting approach, leveraging instantaneous cross-dependencies among meters to enhance data recovery efficiency compared to the conventional vertical inpainting method. Experimental results suggest that the LDMM outperforms other methods, in terms of R2, including the one based on KNN, which yielded the best results in experiments done by Lee & Son (2024). On the flip side, the LDMM has a much higher computational burden, due to its iterative nature, and the matrix inversion computed in each iteration. Therefore, a tradeoff between performance and cost should be considered before choosing the LDMM as a method to fill gaps in power measurements.

7. References

Denhard A, Bandyopadhyay S, Habte A, Sengupta M. Evaluation of Time-Series Gap-Filling Methods for Solar Irradiance Applications. Golden, CO (United States): 2021. https://doi.org/10.2172/1826664.

Iwueze IS, Nwogu EC, Nlebedim VU, Nwosu UI, Chinyem UE. Comparison of Methods of Estimating Missing Values in Time Series. Open J Stat 2018;08:390–9. https://doi.org/10.4236/ojs.2018.82025.

Lee DS, Son SY. PV Forecasting Model Development and Impact Assessment via Imputation of Missing PV Power Data. IEEE Access 2024;12:12843–52. https://doi.org/10.1109/ACCESS.2024.3352038.

Montalvão J, Bastos G, Filho IS. A Simple U-Diffusion Inpainting Structure. TechRxiv 2024. https://doi.org/10.36227/techrxiv.171779330.03561240/v1.