Univariate Time Series Imputation in Industrial Production using Dynamic Harmonic Regression

Viet Ha
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—IconPro is a software company in Aachen providing Industrial AI solutions for predictive quality, predictive maintenance. The company is developing TSAF: Time Series Analysis & Forecasting, an automated, easily integrated module in industrial production workflows or dashboards. A common problem that TSAF has to deal with is a large gap or multiple gaps in the data. Data are univariate time series with or without seasonalities, and are not necessary to have clear trends or any patterns. This thesis uses Dynamic Harmonic Regression (DHR) to impute gaps and evaluation the results in cases of different missing mechanism. The thesis suggests an adaptation of DHR for interpolation, and compares both interpolation and extrapolation DHR imputation to various methods. The suggested interpolation is suitable for Missing At Random (MAR) data. It is consistently the best imputation method for any missing rate among methods under consideration. For Missing Completely At Random (MCAR) data, DHR is not particularly suitable. Competitive methods are ARIMA with a Kalman Filter in case of MCAR, and SARIMA extrapolation in case of MAR. However, DHR interpolation is still preferable as MAR is the more common in reality, and the electricity load dataset is the main interest of the company. The suggested method's positive result is due to the careful handle of seasonalities, first by Fourier terms then by STL decomposition. The autocorrelation of past values is also captured by an ARIMA, so that improves interpolation.**

***Index Terms*—imputation, dynamic harmonic regression, univariate, MCAR, MAR, Fourier terms, multi-seasonality**

I. Introduction

There are many methods developed for multivariate time series imputation, however, univariate time series imputation remains an area of research with not as prolific results, partly because such series do not have inter-attribute correlations to estimate values for missing data [5]. Imputation methods can be classified into simple imputation (non-parametric) and model-based imputation (parametric) [20].

Some simple imputation methods are: LOCF (Last Observation Carried Forward), BOCF (Baseline Observation Carried Forward), NOCB (Next Observation Carried Backward) [1], mean/median/mode imputation, interpolation imputation and moving average (MA) imputation. Model-based imputation [2] includes: spline interpolation, time series regression, autoregressive (AR) model, ARIMA model, dynamic regression model.

This thesis was prepared in partial fulfilment of the requirements for the Degree of Bachelor of Science in Data Science and Artificial Intelligence, Maastricht University. Supervisor(s): Joël Karel, Ralf Peeters

A. ARIMA process and Dynamic Harmonic Regression

AR models that capture autocorrelation are similar to linear regression models, except that the regressors are the past values of the variable [7]. The more general class of AR models is called ARIMA (AutoRegressive Integrated Moving Average). ARIMA models consider not only the autocorrelation of the series values but also the autocorrelation of the imputation errors, in other words, ARIMA combines the autoregressive and moving average model [8]. In addition to information from past observations of a series, Dynamic regression extends ARIMA by including additional information that may be relevant to improve regression imputation. Dynamic Harmonic Regression (DHR) extends further to deal with long seasonal periods or several seasonal periods by making use of Fourier terms of different frequencies [3], so that the seasonalities of series have better approximation.

B. Decomposition and time series components

Time series components are Seasonality, Trend-Cycle and Remainder. A time series can be decomposed into either an additive or multiplicative model [3]. Let S_t, T_t, R_t represent Seasonal, Trend-Cycle, Remainder components, respectively.

- Additive: $Y_t = S_t + T_t + R_t$
- Multiplicative: $Y_t = S_t \times T_t \times R_t$

In additive decomposition, the magnitude of the seasonal component, or the variation around the trend-cycle, does not vary with the change of time (homoscedasticity). Meanwhile, a multiplicative decomposition model is more appropriate when a series has the variation around the trend-cycle is proportional to the level of the time series (heteroscedasticity). Sometimes, complex components make a time series more difficult to impute. The challenging cases are: (1) Complex Seasonality: many seasonalities in the series, (2) Trend-Cycle and Seasonality: long seasonal periods in the series (one can argue long seasonalities are cycles), (3) Heteroscedastic Seasonality: magnitude of the seasonal component varies with the level of the time series.

C. Missing data mechanisms

Rubin [4] classified missing data into 3 categories, according to the mechanism that governs missing: they are Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). Thus, the distribution

of missing data is different depending on what causes the missing data.

In MCAR, no systematic mechanism governs missing data [6]. The cause of missing is uncorrelated to the data. In univariate time series, the probability for a certain observation being missing does not depend on any observed variables, including time, of the series [5]. For example, in electricity load datasets, data points are missing simply because of bad luck, or due to some unknown reasons.

The data missing apart from the loss of information makes MCAR convenient to simulate, however, it does not often reflect reality since missing often occurs due to a reason. If the probability of having a missing value depends on a variable of series, then the data is Missing At Random. In univariate time series - missing depends on the point of time of the observation in the series [5]. For example, data is missing only after a point of time - certain months or periods in a year: the electricity consumption in an area is missing after some weeks after a natural disaster. If the missing is known to be at a certain period and is MCAR, then the data is MAR as well: some uniformly distributed gaps in September.

If the missing values are neither MCAR nor MAR, they are not missing at random (NMAR). The probability of having a missing value depends on reasons that are not observed [22].

D. Literature review and State-of-the-art

There are papers that consider various single imputation methods [9], and evaluate the use of model-based methods such as ARIMA and SARIMA [10].

Jones [26] proposes a method to impute missing observations for stationary series in the Markovian representation of ARMA process. Kohn and Ansley [27] extend the method to interpolate non-stationary series for ARIMA model.

Taylor [18] discusses Multiplicative double seasonal ARIMA model (Box et al. [8] introduce and suggest that can be extended for multiple seasonalities), Holt-Winter exponential smoothing [24], and introduces double seasonal Holt-Winter method. Taylor comments that the introduced method outperforms the other two methods in discussion. Young et al. [11], Harvey and Koopman [25] use a structural approach, put the series into state space form so that Kalman filter and associated recursive algorithms are employed.

De Livera et al. introduce TBATS [19], a generalized framework of exponential smoothing models for series that has both multiple seasonalities and high frequency seasonality. TBATS stands for Trigonometric (to model seasonalities), Box-Cox transformation, ARMA errors, Trend and Seasonal components.

E. Problem statement and research questions

There are also discussions about using the DHR to forecast a series [11] [3] (chapter 10). However, no literature on the use of DHR to impute so far.

The thesis uses DHR to impute the complex case: long seasonal periods and multiple seasonalities in the series. This special kind of series occurs frequently when IconPro deals with industrial electricity datasets, as the sensors record data hourly or at an even shorter time interval. The experiments compare DHR performance with single imputation methods and model-based methods in selected time series, in both MCAR and MAR. Finally, visualization and discussion will be provided.

1) Problem statement: Filling the gaps for univariate time series, especially, the industrial electricity datasets with long seasonal periods and multiple seasonalities in the series.

2) Research questions :

- Research question 1: The complex case is missing values occur in the time series with long seasonal periods and multiple seasonalities. How can DHR be used for imputation in that case VI-A?
- Research question 2: How well does DHR perform in cases where there are long seasonal periods and multiple seasonalities VI-A?
- Research question 3: How well does DHR perform in cases where the data is missing at MCAR and MAR VI-A?

II. Imputation using DHR and other methods

Imputation from model-based methods in this thesis uses prediction from regression. The idea is to reconstruct a time series by forecasting from its available values. There are concerns for this imputation technique, will be further examined in discussion.

DHR may refer to: DHR with Time Variable Parameters (DHRTVP) [11] and DHR with Fourier terms (DHRFT) [3]. The technique employed in this thesis is the latter.

A. Handling Trend-Cycle and Seasonal using DHRFT

Trends can be categorized into deterministic and stochastic trends. DHR can capture deterministic trends using normal regression, where the disturbance follows a ARMA process ($d = 0$); or stochastic trends where the disturbance follows a ARIMA process ($d = 1$), d is the difference order [3].

Let s be the number of seasons in the data. To model seasonal effects γ_t using dummy variables: $\gamma_1, \gamma_2, ...\gamma_s$ represents seasonal values and $\Sigma_{i=1}^{s}\gamma_t = 0$. The effect summed over the seasons should equal zero: $\gamma_{t+1} = -\Sigma_{j=1}^{s-1}\gamma_{t+1-j}$. To allow the pattern to change over time, introduce a disturbance term: $\gamma_{t+1} = -\Sigma_{i=1}^{s-1}\gamma_{t+1-j} + \omega_t, \omega_t \sim N(0, \sigma^2)$ [12] (section 3.2.2). The expectation of the sum of the seasonal effects is zero. The seasonal pattern is eliminated if the series is aggregated over s consecutive time periods, thus there is a separation of trend and the seasonal. As an alternative for γ_t representation, one can use Trigonometric terms of seasonal frequencies,$\gamma_t = \Sigma_{j=1}^{[s/2]}\alpha_j cos(\lambda_j t) + \beta_j sin(\lambda_j t)$ where α_j, β_j are trigonometric parameters, $[s/2]$ is s if s is even and $(s-1)/2$ if s is odd, and $\lambda_j = 2\pi j/s$, $j = 1, ..., [s/2]$ [13] (section 5.5).

Therefore, seasonal trigonometric representation allows DHRFT to capture both Trend-Cycle and Seasonal components.

B. DHR with Fourier terms and DHR with Time-Varying Parameters (TVPs)

1) The DHR with TVPs model (DHRTVP): contains the trend, cyclical, seasonal and white noise component, $(T_t + C_t + S_t + e_t)$, from the general form of Unobserved Component Model (UCM) [13] as below:

$Y_t = (T_t + C_t + S_t + e_t) + f(u_t) + N_t$, $e_t \sim N(0, \sigma^2)$ [11]

where Y_t is the observed time series, T_t, C_t, S_t, $f(u_t)$, N_t, e_t represent trend, (quasi-)cyclical, seasonal, impact of external information u_t, stochastic disturbance model, white noise components, respectively.

The Seasonal and Cyclical are modelled as

$S_t = \Sigma_{i=1}^{R_s} a_{i,t} cos(\omega_i t) + b_{i,t} sin(\omega_i t)$

$C_t = \Sigma_{i=1}^{R_c} \alpha_{i,t} cos(f_i t) + \beta_{i,t} sin(f_i t)$

where $a_{i,t}, b_{i,t}, \alpha_{i,t}, \beta_{i,t}$ are non-stationary stochastic variables TVPs; $\omega_i, i = 1, .., R_s$, $f_i, i = 1, .., R_c$ are frequencies associated with seasonality and cyclical component, respectively [11]. Components are stochastic in DHRTVP.

2) The DHR with Fourier terms model (DHRFT): has more focus on a component, N_t. The stochastic disturbance is handled using ARIMA models.

$Y_t = T_t + C_t + S_t + N_t + e_t$, $e_t \sim N(0, \sigma^2)$

The components C_t, S_t are deterministic components, because the parameters $a_{i,t}, b_{i,t}, \alpha_{i,t}, \beta_{i,t}$ are no longer stochastic. Trend T_t in this model is also deterministic.

DHRFT is often written in regression form where $x_{i,t}$ is a Fourier term, if η_t follows ARIMA(1, 1, 1)

$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \eta_t$

$\eta_t - \eta_{t-1} - d = \phi_1(\eta_{t-1} - \eta_{t-2} - d) + \epsilon_t + \theta_1 \epsilon_{t-1}$ [3]

in which: d is the drift term, ϕ_1 is the parameter of $(\eta_{t-1} - \eta_{t-2} - d)$, θ_1 is the parameter of ϵ_{t-1}, d is η_t is the errors of regression (stochastic disturbance N_t in UC model), ϵ_t is a white noise series - the errors of ARIMA [3].

C. Methods in consideration

This section provides a more detailed explanation of DHRFT and overview of competitive methods in comparison.

1) Simple imputation: Mean imputation, LOCF, linear interpolation.

2) Regression models with deterministic trends: The observed series (Y_t) is represented as $Y_t = \mu_t + \epsilon_t$ [28] where μ_t is a deterministic function that is periodic (explained part of the model) and the unobserved variation (ϵ_t) around μ_t (random 'error') has zero mean for all t, $E(\epsilon_t) = 0 \ \forall t$. The assumptions about errors in these models are: they are not autocorrelated, and they are uncorrelated to the systematic part of the model.

- Regression with deterministic seasonal trend (seasonal means model): The k-period seasonal trend contains k parameters β_i $i = 1, k$, where $\mu_t = \beta_i$ for $t = i, i+k, i+2k, ...$ The regressors x_i are seasonal dummy variables [28].
- Regression with cosine trend: A regression using Fourier terms as regressors can capture seasonalities. In the simplest model with only k-period seasonality, $\mu_t = \beta_0 + \beta_1 cos(2\pi t/m) + \beta_1 sin(2\pi t/m)$ [28].

3) ARIMA models: ARIMA is briefly described above. Detailed models description as below.

- The autoregressive model, AR(p), can be written as a multiple regression: $Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t$ where regressors are lagged values of Y_t and ϵ_t is white noise [8].
- The moving average model, MA(q), contains previous imputation errors as regressors:$Y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_p \epsilon_{t-p}$ where ϵ_t is white noise [8].
- SARIMA: In addition to ARIMA terms, values at seasonal lags are also included in SARIMA: ARIMA$(p, d, q)(P, D, Q)_m$ where m is the seasonal periods [3].
- STL+ARIMA: STL method decomposes a time series into Trend, Seasonal and Remainder components using Loess. Use ARIMA to generate forecasts for the de-seasonalized time series. The results are then re-seasonalized by adding a seasonal naive forecast to provide imputation [14] [15].

4) Dynamic Harmonic Regression: In dynamic regression (DR), the error terms η_t are allowed to be autocorrelated, unlike in the normal regression where errors ϵ_t are white noise series, so one can say, DR is the regression with ARIMA errors.

- Trend: In DR, linear trend is considered as following regression, $Y_t = \beta_0 + \beta_1 t + \eta_t$. Trend is either deterministic where the disturbance follows a ARMA process, or stochastic where the disturbance follows a ARIMA process [3]
- Seasonal and Cyclical components: a DR with Fourier terms (DHRFT) of different frequencies can capture the data with more than one frequency.

The trend component T_t can be included into the Cyclical or Seasonal component as a zero-frequency term [11]. The Cyclical and Seasonal are modelled in a same manner using series of cosine and sine terms, so the complete DHRTVP can be written as:

$Y_t = \Sigma_{i=0}^{R} s_t^{p_i} + \epsilon_t$, $\epsilon_t \sim N(0, \sigma^2)$

$Y_t = \Sigma_{i=0}^{R} [a_{i,t} cos(\omega_i t) + \Sigma_{i=1}^{n} b_{i,t} sin(\omega_i t)] + \epsilon_t$ [16]

where $\omega_i = 2\pi f_i, i = 1, .., R$ are the fundamental and harmonic frequencies of the sinusoidal components associated with the ith component, and R oscillatory components are estimated using AR model [11].

Therefore, DHRTVP is an extension of the constant parameters DHRFT below.

$Y_t = a + bt + \Sigma_{k=1}^{K} [\alpha_k s_k(t) + \Sigma_{i=1}^{n} \beta_k c_k(t)] + \Sigma_{i=1}^{p} \phi_i Y_{t-i} + \Sigma_{i=1}^{q} \theta_i \epsilon_{t-q} + \epsilon_t$, $\epsilon_t \sim N(0, \sigma^2)$ [3]

where linear global trend $T_t = a + bt$, and seasonal component $S_t = \Sigma_{k=1}^{K} [\alpha_k s_k(t) + \Sigma_{i=1}^{n} \beta_k c_k(t)]$ as a sum of sine and cosine terms: $s_k = sin(2\pi kt/m), c_k = cos(2\pi kt/m)$, [3] and stochastic disturbance $N_t = \Sigma_{i=1}^{p} \phi_i Y_{t-i} + \Sigma_{i=1}^{q} \theta_i \epsilon_{t-q}$ as a ARMA(p, q) process.

D. Time series regression, ARIMA, DHRFT: similarities, differences and generalization

This section provides a theoretical comparison between DHRFT and competitive methods. Then, a general framework is used to better show similarities, differences and relations between models.

1) DHRFT and time series regression with cosine trend: Both models capture seasonalities and cycles using Fourier terms as regressors, but subtle short-time time series dynamics are only handled in DHRFT using ARMA process.

2) DHRFT and (S)ARIMA: Both models include information from the past values. DHRFT uses Fourier terms to capture Seasonal and Trend-Cycle components for regression, SARIMA captures seasonalities using additional information from seasonal lags.

3) Generalization: ARIMA, DHR are special cases of the Unobserved Component Model (UCM), multiple regression with time-varying coefficients, first introduced by A.C.Harvey [13]:

$Y_t = T_t + C_t + S_t + f(u_t) + N_t + \epsilon_t$, $\epsilon_t \sim N(0, \sigma^2)$

in which, $f(u_t)$ allows one to include the impact of external information through vector u_t, and N_t is a stochastic disturbance model [11].

In UCM framework, the methods in consideration:

- Time series regression with cosine trend
  $Y_t = T_t + C_t + S_t + \epsilon_t$, trend is modelled as constant parameter, harmonic regression.
- ARIMA (Box-Jenkins model)
  $Y_t = T_t + C_t + S_t + N_t + \epsilon_t$, use autoregressive component representation to capture unobserved components, ARIMA has another form of State Space Model.
- DHR with TVPs
  $Y_t = T_t + C_t + S_t + \epsilon_t$ where components are characterized by stochastic, TVPs allows N_t included in $T_t + C_t + S_t$.
- DHR with Fourier terms
  $Y_t = T_t + C_t + S_t + N_t + \epsilon_t$ = Time series regression with Fourier terms (Seasonal) + ARIMA (Dynamics: $D_t = T_t + C_t + N_t + \epsilon$)

E. Estimation of parameters in DHRFT

Estimated parameters in DHRFT are coefficients of the regression model that minimize the sum of squared errors ϵ_t. This is the error from ARMA model, not the error from regression model η_t. Minimization of the sum of squared η_t, means ignoring the autocorrelations in the errors, which will lead to undesirable results [17]. An alternation is Maximum Likelihood Estimation, which gives similar estimates.

F. DHRFT adaptation for imputation

Because DHRFT is initially designed to forecast a time series, a natural idea for imputation is to use the extrapolation result. Although this approach seems so simple at first, it makes sense, especially in case of series with MCAR. The reason is, gaps in such series usually are quite small, thus, the trend underneath tends to follow the trend of nearby values. That means, in such cases, the short-term extrapolation is comparatively good to interpolation for small gaps imputation, given the well-handled seasonalities methods.

In case of MAR, often a large gap occurs, such extrapolation is questionable, because forecasting is rarely accurate in long-term as the later forecast based on earlier incorrect forecast values, thus, errors accumulate. This section presents an adaptive algorithm for imputation in such cases.

Step 1: Capture the seasonal component of the series using Regression with Fourier terms, then obtain imputation, S_{imp}, for seasonal by prediction from the regression. The residual is a Dynamic component, $D_t = T_t + C_t + N_t + \epsilon$, which contains rich information, actually, all information excepts the partly captured seasonalities.

Step 2: Obtain the seasonally adjusted of Dynamic component, $D'_t = D_t - S_d$, using STL decomposition to fit an ARIMA model.

Step 3: Obtain the trend of Dynamics, T_d, using Kalman filter and smoother for ARIMA. Re-seasonalizing by adding back seasonal component of Dynamics, S_d, (obtained from STL in step 2) to get imputed Dynamic component, $D_{imp} = T_d + S_d$, for original series.

Step 4: Add imputed Seasonal component to imputed Dynamic component to get the imputation, $Y_{imp} = S_{imp} + D_{imp}$, for original series.

III. Experiments

A. Dataset

1) Dataset Taylor: Data were collected at intervals of 30 minutes from 05-06-2000 to 27-08-2000 for electricity demand in England and Wales. Units: Megawatts [18]. The dataset is chosen to expose methods to a time series with complex seasonality. In addition, the cycle component has a very low frequency.

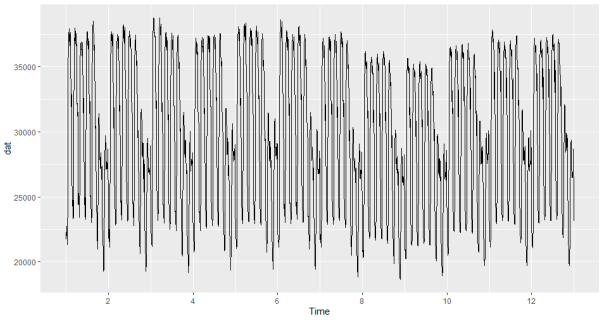

Fig. 1. Taylor: Half-hourly electricity demand in England and Wales from Monday 5 June 2000 to Sunday 27 August 2000

Dataset represents the challenging case, where DHRFT may give appropriate results. Description in the table below.

Trend-Cycle	Seasonal	Remainder
Non-linear Trend	Complex	Homoscedastic
12 Weekly Cycles 336=48*7 periods	Long: 48 periods Long: 336 periods	

Autocorrelation plot shows that data has strong positive and negative autocorrelation, a typical pattern for seasonality.

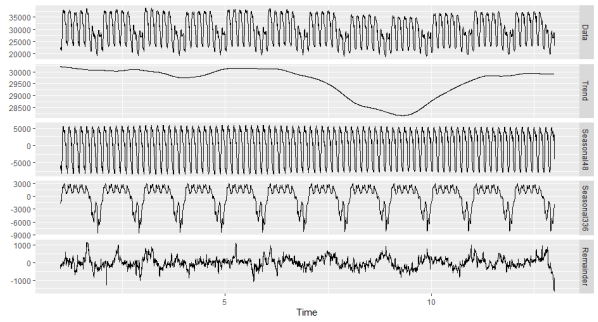

Fig. 2. Components of Taylor dataset

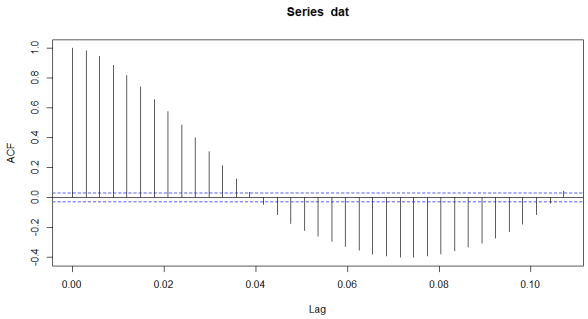

Fig. 3. Autocorrelation plot for Taylor dataset

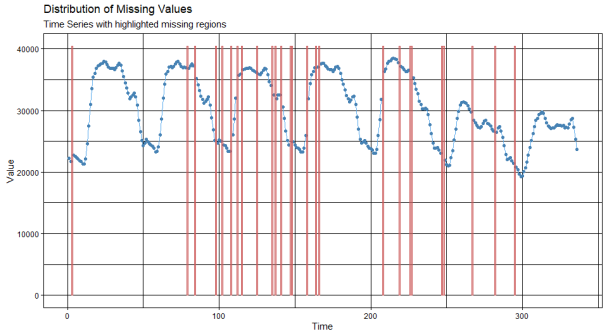

Fig. 4. Distribution of Missing Values (MCAR, Missing Rate: 10%)

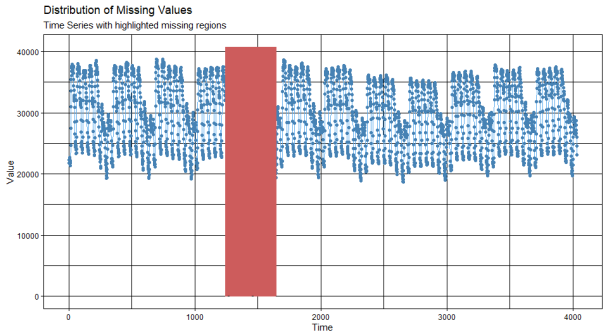

Fig. 5. Distribution of Missing Values (MAR, Missing rate: 10%)

2) Dataset Calls: Data were collected at intervals of five minutes for the call volume processed on weekdays from 7:00 am to 9:05 pm at a large North American commercial bank [19]. Dataset represents the challenging case, where DHRFT may give appropriate results. It is chosen to confirm if the "successful methods" for the Taylor dataset are repeatable for similar complex dataset with even longer seasonality, or are merely lucky. Description in the table below.

Trend-Cycle	Seasonal	Remainder
Non-linear Trend	Complex	Homoscedastic
	Long: 169 periods Long: 845 periods	

B. Missing data simulation

Data missing mechanisms are described above, experiments simulate MCAR and MAR. For univariate series, r is the event that an observation being missing,

- MCAR: $P(r|Y_{observed}, Y_{missing}) = P(r)$
- MAR: $P(r|Y_{observed}, Y_{missing}) = P(r|Y_{observed})$ [5]

First, a missing rate is set, then missing data are taken out from a complete time series in below manners.

- MCAR: Generate values within interval [0, 1] from a continuous uniform distribution. Any values less than the missing rate are set to NA.
- MAR: Randomly select a time point in the series and make a gap with length corresponding to the missing rate.

C. Experiments plan

The experiments plan includes the following steps:

- Start with a complete time series ts_{comp}.
- Remove data points according to missing mechanism (MCAR, MAR) to create an incomplete time series ts_{miss}.
- Apply imputation methods to ts_{miss} to get ts_{imp}.
- Compare ts_{imp} to ts_{comp}.

Adjustable variables in the experiments:

- Random seed (RS) for initial values of the start index of gap (MAR), or indexes of missing (MCAR).
- Missing rate (MR) to determine the size of gap (MAR), or how many percentages of missing values in the series (MCAR).

D. Evaluation

Metrics in consideration to measure imputation perform are:

- Forecast error: The difference between the actual value and its prediction based on all past observations, $e_t = y_{(t)} - \hat{y}_{(t|t-1)}$ where $y_{(t)}$ is the observation at time t, $\hat{y}_{(t|t-1)}$ denotes the prediction of $y_{(t)}$ based on all past observations.
- MAE (Mean Absolute Error): $MAE = mean(|e_t|)$
- RMSE (Root Mean Square Error): $RSME = \sqrt{mean(e_t^2)}$
- MAPE (Mean Absolute Percentage Error): $MAPE = mean(|p_t|)$ where $p_t = 100e_t/y_t$ is the percentage error.

MAE and RMSE are scale-dependent. A benefit is the unit of these metrics is on the same scale of the series, thus, is easily understandable. They are also chosen to overcome the disadvantage of MAPE (undefined when $y_t = 0$ or very large when y_t is close to zero). Both are widely used, and are known in papers as L1 loss and L2 loss.

MAPE is chosen to communicate the results to electrical engineers and energy managers, the end users of the research, due to its ease of understandability. MAPE does not have unit, thus, it is possible to compare accuracy between datasets.

Interpretations of MAPE values in forecasting, according to Lewis [23], are adapted to interpret imputation accuracy.

MAPE	Interpretation
$<$ 10%	High degree of accuracy imputation
10% - 20%	Good accuracy rate imputation
20% - 50%	Reasonable imputation
$>$ 50%	Inaccurate imputation

IV. Results

- Extrapolation: Imputation using the forecast from the model, fitted by data before gap with the length of at least 3 longest seasonal.
- Interpolation: The model is fitted by data from both sides of the gap, then a Kalman filter is used to fill the gap.

A. Imputation in case of MCAR

DHRFT uses the Fourier terms of order: 10 and 15. Reference tables and figures are: I, II, III, IV, V, VI and 6, 14, 15 Key observations from the results

As can be seen from Figure 6, 14, 15, among the methods, linear interpolation comes second, and STL + linear interpolation is the best method.

To see if a model-based method improves accuracy, one only need to compare that method to linear interpolation (STL+interpolation is not considered here, as STL is an algorithm that deals with seasonalities, hence, not really a simple method), because it outperforms other simple methods at any missing rate. Results from Table I suggest that model-based methods (ARIMA, DHRFT, BSM) outperform simple imputation methods.

Using the MAPE interpretation III-D above and Table VI, Table I, when the missing rate is less than 10%, all model-based methods, and linear interpolation provide high degree of accuracy imputation.

Results of Table II, when the missing rate is between 10% and 30%, all model-based methods provide good accuracy rate imputation. ARIMA, BSM performance are slightly better than DHRFT.

Results from Table III, IV, V show that when the missing rate higher than 50%, all methods under consideration provide inaccurate imputation. BSM outperforms the other model-based methods.

B. Imputation in case of MAR

1) Taylor dataset: Reference tables are: VII, VIII, IX, X. Key observations from the results:

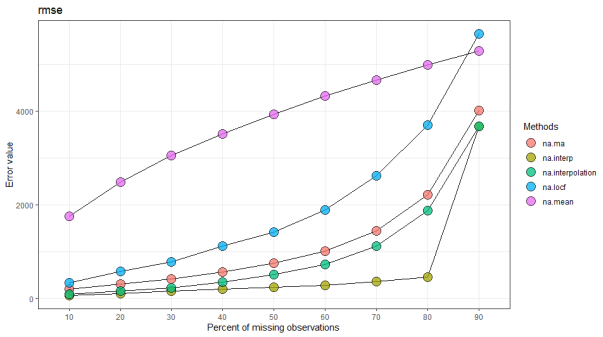

Fig. 6. Simple imputation methods performance

TABLE I
IMPUTATION METHODS PERFORMANCE FOR MCAR, MISSING RATE: 10%, RANDOM SEED: 10 AND 100*

Dataset Taylor	RMSE	MAE	MAPE
Mean	1781.848	496.504	1.803138
	1684.3*	451.9869*	1.614654*
LOCF	415.1887	82.22049	0.2861251
	310.5561*	64.38343*	0.2250505*
Linear Interpolation	88.35853	19.93262	0.06962146
	73.6909*	16.09135*	0.05768408*
ARIMA + Kalman	55.16056	**11.77723**	**0.04134448**
	72.52243*	11.23335*	0.04033905*
DHRFT + Kalman	**54.80087**	12.81477	0.04505768
	46.53528*	**11.03915***	**0.03915016***
BSM + Kalman	56.2795	12.13909	0.04266873
	48.19006*	11.35807*	0.04051567*

TABLE II
IMPUTATION METHODS PERFORMANCE FOR MCAR, MISSING RATE: 30%, RANDOM SEED: 10 AND 100*

Dataset Taylor	RMSE	MAE	MAPE
ARIMA + Kalman	154.996	**48.87798**	**0.1733423**
	137.6643*	**45.29473***	**0.1586055***
DHRFT + Kalman	158.2965	53.63028	0.1873732
	150.4457*	51.109*	0.1794831*
BSM + Kalman	**144.9342**	50.55921	0.1783337
	128.2535*	47.23132*	0.166212*

TABLE III
IMPUTATION METHODS PERFORMANCE FOR MCAR, MISSING RATE: 50%, RANDOM SEED: 10 AND 100*

Dataset Taylor	RMSE	MAE	MAPE
ARIMA + Kalman	407.6029	156.8031	**0.5474221**
	425.7453*	**155.8839***	**0.5487401***
DHRFT + Kalman	467.2726	12.81477	0.6333252
	460.8112*	181.8359*	0.6538035*
BSM + Kalman	**390.9527**	**156.2406**	0.5476756
	468.2507*	160.2425*	0.5701367*

TABLE IV
IMPUTATION METHODS PERFORMANCE FOR MCAR, MISSING RATE: 70%, RANDOM SEED: 10 AND 100*

Dataset Taylor	RMSE	MAE	MAPE
ARIMA + Kalman	1786.478	626.6449	2.204738
	1145.794*	599.9613*	2.092932*
DHRFT + Kalman	**1159.154**	513.9245	1.853952
	999.097*	509.6667*	1.815796*
BSM + Kalman	1279.521	**497.9766**	**1.791056**
	880.0917*	**450.0473***	**1.587116***

TABLE V
IMPUTATION METHODS PERFORMANCE FOR MCAR, MISSING RATE: 90%, RANDOM SEED: 10 AND 100*

Dataset Taylor	RMSE	MAE	MAPE
ARIMA + Kalman	13678.47	7848.759	26.87326
	7272.434*	4578.69*	15.07571*
DHRFT + Kalman	NA	NA	NA
	NA*	NA*	NA*
BSM + Kalman	**3261.75**	**2038.147**	**7.420709**
	3306.665*	**2071.685***	**7.51738***

TABLE VI
IMPUTATION METHODS PERFORMANCE FOR MCAR, MISSING RATE: 5%, RANDOM SEED: 10 AND 100*

Dataset Taylor	RMSE	MAE	MAPE
ARIMA + Kalman	35.08807	**6.046277**	0.02131483
	30.87681*	**5.232663***	**0.01852103***
DHRFT + Kalman	**34.63321**	6.049175	**0.02130214**
	31.54522*	5.32484*	0.01894877*
BSM + Kalman	36.75876	6.181347	0.02199929
	34.76878*	5.758841*	0.02059875*

- DHRFT interpolation outperforms other methods in any rate.
- Regression (dummy seasonal, cosine trend, Fourier terms) alone cannot fully capture seasonality.
- Best performers (DHRFT, STL+ARIMA, SARIMA) include ARIMA process to capture seasonality left in the dynamics.
- SARIMA and STL+ARIMA are the best methods among extrapolation.

As can be seen from Table VII, VIII, when the missing rate is less than 10%, methods using ARIMA provide good accuracy rate imputation.

Results from Table IX show that none of the methods under consideration provide good accuracy when the missing rate is 20%. They provide reasonable imputation.

Results from Table X show that all methods provide inaccurate imputation when the missing rate is (and is higher than) 30%.

TABLE VII
IMPUTATION METHODS PERFORMANCE FOR MAR, MISSING RATE: 5%, RANDOM SEED: 11 AND 111*

Dataset Taylor	RMSE	MAE	MAPE
DHRFT, extrapolation	105.6213	19.09429	0.06596858
	124.4202*	22.28138*	0.07317242*
DHRFT, interpolation	**48.26414**	**8.019815**	**0.02767675**
	66.3261*	**12.12653***	**0.04252004***
STL + ARIMA	101.9288	18.41576	0.06337597
	123.0416*	21.87764*	0.07187323*
SARIMA	101.0872	19.31313	0.06787913
	131.6407*	25.62177*	0.08969087*
Regression: dummy seasonal	80.72412	13.12245	0.04569009
	213.9308*	43.63511*	0.1472104*
Regression: cosine trend	410.0039	75.5922	0.265653
	391.4375*	72.60027*	0.2595283*
Regression: Fourier terms	81.57807	13.2749	0.04626492
	213.6456*	43.70113*	0.1474486*

TABLE VIII
IMPUTATION METHODS PERFORMANCE FOR MAR, MISSING RATE: 10%, RANDOM SEED: 11 AND 111*

Dataset Taylor	RMSE	MAE	MAPE
DHRFT, extrapolation	224.0584	57.4834	0.1858772
	180.6532*	46.62774*	0.1655642*
DHRFT, interpolation	**123.856**	**30.69019**	**0.1006513**
	141.753*	**34.62198***	**0.1313986***
STL + ARIMA	207.7226	52.46011	0.1691204
	198.001*	50.38493*	0.176178*
SARIMA	201.6618	53.82574	0.1759446
	202.355*	53.00067*	0.1894844*
Regression: dummy seasonal	189.6209	44.88509	0.1437164
	272.8222*	72.78268*	0.2496814*
Regression: cosine trend	612.4713	156.4037	0.5269943
	617.4531*	162.7505*	0.6022391*
Regression: Fourier terms	189.6398	44.99005	0.1441025
	272.7116*	72.89922*	0.2501697*

TABLE IX
IMPUTATION METHODS PERFORMANCE FOR MAR, MISSING RATE: 20%, RANDOM SEED: 11 AND 111*

Dataset Taylor	RMSE	MAE	MAPE
DHRFT, extrapolation	345.8994	125.962	0.4143229
	402.0623*	145.6519*	0.5194539*
DHRFT, interpolation	**261.8725**	**90.01337**	**0.2834949**
	304.7766*	**106.0591***	**0.371084***
STL + ARIMA	304.0123	105.5171	0.3450599
	321.104*	112.9678*	0.3954806*
SARIMA	295.7822	108.9851	0.3614779
	319.9685*	113.0446*	0.3976359*
Regression: dummy seasonal	305.6867	104.6049	0.3418382
	392.088*	145.9585*	0.505731*
Regression: cosine trend	896.2089	322.3565	1.073523
	873.555*	324.549*	1.188009*
Regression: Fourier terms	305.9289	104.8632	0.3427774
	391.788*	146.0594*	0.5063323*

TABLE X
IMPUTATION METHODS PERFORMANCE FOR MAR, MISSING RATE: 30%, RANDOM SEED: 110 AND 111*

Dataset Taylor	RMSE	MAE	MAPE
DHRFT, extrapolation	636.9223	281.8182	0.9605541
	509.4058*	212.2104*	0.7351961*
DHRFT, interpolation	**361.3602**	**145.599**	**0.5047027**
	435.7641*	**186.5176***	**0.6408963***
STL + ARIMA	631.9802	279.1762	1.001097
	515.7464*	215.3012*	0.7483487*
SARIMA	554.9218	248.187	0.8795605
	482.391*	202.0824*	0.69802*
Regression: dummy seasonal	570.8019	249.0438	0.8700572
	634.9441*	281.9848*	0.9715588*
Regression: cosine trend	1120.041	502.0359	1.80451
	1149.243*	512.7476*	1.841231*
Regression: Fourier terms	570.6511	248.7671	0.8692464
	634.6677*	281.99*	0.971829*

TABLE XI
IMPUTATION METHODS PERFORMANCE FOR MAR, MISSING RATE: 5%, RANDOM SEED: 111 AND 11*

Dataset Calls	RMSE	MAE	MAPE
DHRFT, extrapolation	7.376253	1.328042	0.7219815
	4.475746*	0.7810242*	0.4740498*
DHRFT, interpolation	**4.237058**	**0.744275**	**0.4605721**
	4.596049*	**0.7585797***	**0.4094515***
STL + ARIMA	6.462096	1.108664	0.5776044
	4.671047*	0.8169099*	0.4923291*
Regression: dummy seasonal	7.860741	1.386061	0.7455102
	4.638741*	0.8100706*	0.5007589*
Regression: Fourier terms	7.667592	1.356998	0.7292912
	4.364184*	0.7640647*	0.4784791*

TABLE XII
IMPUTATION METHODS PERFORMANCE FOR MAR, MISSING RATE: 10%, RANDOM SEED: 11 AND 111*

Dataset Calls	RMSE	MAE	MAPE
DHRFT, extrapolation	6.975018	1.662469	0.9760688
	9.387474*	2.312065*	1.273853*
DHRFT, interpolation	7.950382	1.987818	1.218956
	8.033972*	**1.850106***	**0.9552571***
STL + ARIMA	7.670953	1.902922	1.100529
	8.418758*	2.048299*	1.09014*
Regression: dummy seasonal	6.514746	1.601975	0.9625689
	11.80134*	3.139866*	1.781558*
Regression: Fourier terms	**5.967162**	**1.410377**	**0.8596373**
	11.09081*	2.845837*	1.612543*

2) Calls dataset - extrapolation: Reference tables are: XI, XII. Key observations from the results

- Dataset Calls has complex seasonality, and higher periods (169, 845) than Taylor (48, 336), but the other findings stay the same.
- DHRFT interpolation continue to outperform other extrapolation methods at any rate.

Caution: For this dataset, MAPE interpretation is invalid as many zero values in the series. One can only rely on RMSE and MAE to say DHRFT interpolation is better, however, it is uncertain to conclude that DHRFT provides good accuracy imputation.

C. Diagnostic check for Taylor dataset: Residuals plots

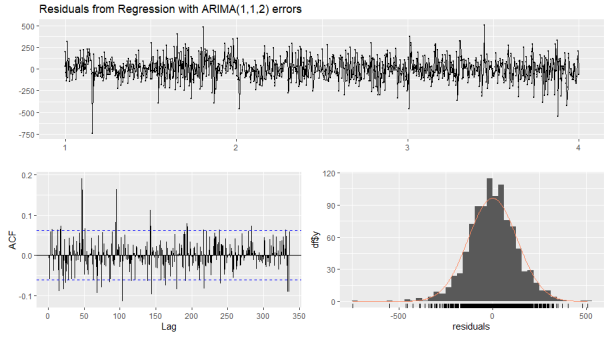

Fig. 7. MAR, 5%: DHR innovation residuals (taylor)

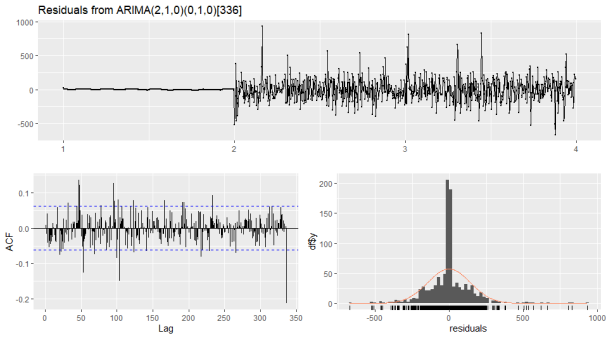

Fig. 8. MAR, 5%: SARIMA innovation residuals (taylor)

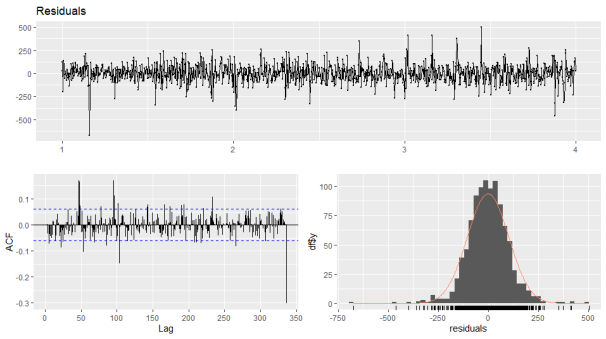

Fig. 9. MAR, 5%: STL+ARIMA innovation residuals (taylor)

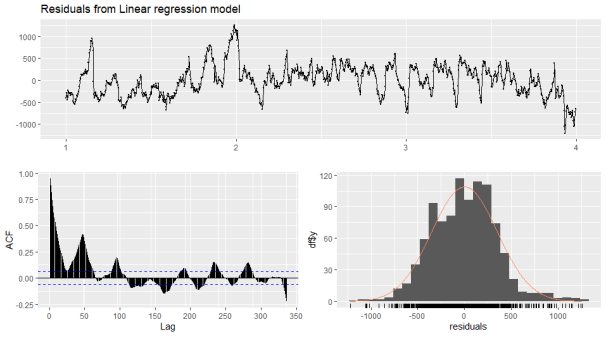

Fig. 10. MAR, 5%: Regression with Dummy Seasonal residuals (taylor)

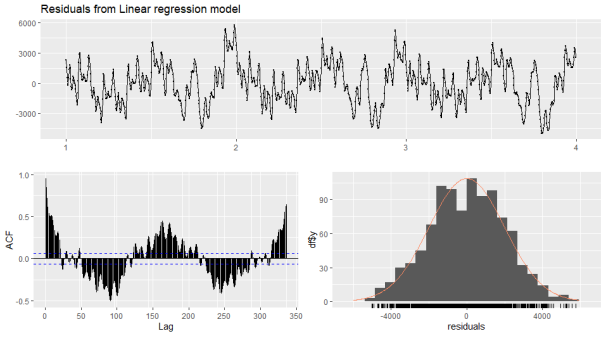

Fig. 11. MAR, 5%: Regression with Cosine trend residuals (taylor)

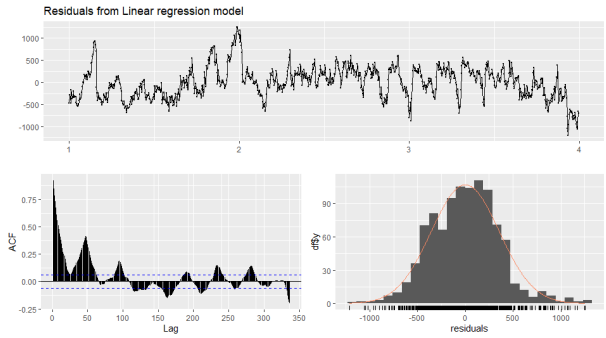

Fig. 12. MAR, 5%: Regression with Fourier terms residuals (taylor)

Key observations from residuals plots: None of the methods is able to completely capture information of autocorrelations (ACF plots for residuals still contain significant spikes).

Figure 8 shows that, SARIMA almost captures the autocorrelation between lags (less significant spikes at nearby values) of the original series. Lags at the period-length time suggests that seasonality in the series is not fully captured. There are many zero values in residual distribution due to the seasonal difference, $d = length(seasonality) = 336$.

As can be seen from Figure 7, DHRFT almost captures the autocorrelation between lags of the harmonic regression errors series. Lags at the period-length time suggests that seasonality of the errors is not fully captured.

As can be seen from Figure 9, STL+ARIMA almost captures the autocorrelation between lags of the ARIMA model fitted by the seasonal adjusted series. The distribution of residuals has the lowest variance among residuals distributions of methods in consideration. This finding is actually the idea behind step 2 of the suggested algorithm, as STL seems to preprocess the seasonalities well. Lags at the period-length time suggests that seasonality of the seasonal adjusted series is not fully captured.

Residuals plots of Regression with Dummy Seasonal, Figure 10, are very alike to residuals plots of Regression with Fourier terms, Figure 12. That suggests the similarity of performance between methods. Both methods do not capture the autocorrelation in the series.

As can be seen from Figure 11: Plots of residuals of Regression with deterministic Cosine trend show that the method fails to capture the autocorrelation in the series. The distribution of residuals has the highest variance.

V. Discussion

The quality of imputation depends on information can be captured by models. Simple methods fail to capture both seasonality and autocorrelations in the series, thus, performance is generally bad. Therefore, it is recommended to avoid these methods for seasonal and cyclical series.

A. Metrics discussion

Many metrics are chosen so that they compensate each other disadvantages.

MAE and RMSE are on the same scale of the series, hence, are easy to understand. MAE is easy to interpret and is useful when compare methods in a single time series. Both MAE and RMSE cannot be used to compare method performance across series with different units. On the other hand, MAPE has the advantage of ability of comparisons between datasets. However, a percentage based measure like MAPE, at some time t, is undefined if the actual value is zero, $y_t = 0$, or is extremely large if the actual value, y_t, is close to zero [3]. Another issue is, MAPE makes no sense when the unit of measurement does not have a meaningful zero. This is not the case in electricity datasets, as zero Megawatts represents no consumption. In this dataset, MAPE is reliable as zero values are removed, and it is unlikely that people do not use electricity at all. For the Calls dataset, zero value, $y_t = 0$, means there is no call at that time t. In this dataset, there are many near zero value, thus, the use of MAPE is not recommended, because it wrongly magnifies errors.

B. Imputation in case of MCAR

DHRFT + Kalman: DHR handles non-stationary and seasonality through regression terms (Fourier terms). By checking the ACF plot of residuals, this model, however, fails to capture the seasonality shown in the significant spikes at lags 48, 96 corresponding to daily seasonal 7. Some additional experiments show that adding more Fourier terms with different frequencies or increasing Fourier orders does not improve the quality of imputation. This recommends DHRFT alone is not enough to fully handle complex seasonality.

ARIMA + Kalman: Checking the residuals shows that ARIMA fails to capture the seasonality. There are significant spikes at lags 14, 24, 33, 34, 9 which seems really unexplainable. However, the model tries harder to include the information from recent past - higher order for MA part. This is meaningful given the dataset of half-hourly electricity load, expressed differently, the demand for a specific time is more likely similar to the demand of nearby previous hours than the demand of the same time yesterday. That explains the competitive good results that ARIMA provides.

Basic Structural Model + Kalman: This model allows both components stochastic, that means trend are local, with time-varying slope in the dynamics for underlying level μ_t, and seasonal component with dynamics. The model [29] is given by:
Measurement: $x_t = \mu_t + \gamma_t + \epsilon_t,\ \epsilon_t \sim N(0, \sigma^2_{\epsilon_t})$
Local Trend: $\mu_{t+1} = \mu_t + \nu_t + \xi_t,\ \xi_t \sim N(0, \sigma^2_{\xi_t})$
Slope: $\nu_{t+1} = \nu_t + \zeta_t,\ \zeta_t \sim N(0, \sigma^2_{\zeta_t})$
Seasonality: $\gamma_t = -\gamma_t - ... - \gamma_{t-s+2} + \omega_t,\ \omega \sim N(0, \sigma^2_{\omega_t})$
The stochastic components allow adaptability, especially, imputation improvement is shown for high missing rate, meanwhile, there is little difference for low missing rate.

The experiments expose some limitations of DHRFT methods. The first issue is not really a weakness of DHRFT, nevertheless, it's practical to question the seasonality effect in comparison with autocorrelation effect of nearby value. It's not only useful for interpretation, but also for suitable

models selection. For electricity load dataset, a non-seasonal model can give competitive results to DHRFT, just by simply including more lags from the past. Second, DHRFT models a stochastic seasonality component using a deterministic function of time, which is undesirable, and handling the dynamics by ARIMA. This approach assumes fixed seasonality, as the parameters of regression do not allow time-varying. DHRTVP, however, allows changes in seasonality and should be preferable. Third, although DHRFT explicitly models the seasonality, its dynamics, again, shows seasonal pattern. That means, regression with Fourier terms alone cannot capture all seasonality information, no matter how many more terms added, so the performance depends on how well the ARIMA models handle dynamics' seasonalities. This explains, DHRFT cannot always outperform a good normal ARIMA model. After all, ARIMA for dynamics may not be suitable because the remainder part is non-stationary, in such case, a normal ARIMA is preferable to DHRFT.

C. Imputation in case of MAR

1) Interpolation and extrapolation: In general, interpolation methods use information from both ends of the gap of the series to impute, thus, outperforms extrapolation methods which use only information from the left side of the gap.

2) Dataset Taylor: The trend underneath of this dataset does not vary drastically, (electricity load through weeks does not easily change in short time), is an implicit assumption of imputation. It allows the well-handling seasonal component methods to shine, as the trend interpolation does not greatly affect the result. The suggested DHRFT adaption is the best method for any rate because it tries harder to improve seasonal component. In the first step, seasonalities are initially captured by Fourier terms such that the mean squared error η_t^2 is minimalized. That means to minimize the dynamics, $D_t = T_t + C_t + N_t + \epsilon$, this component, itself, contains information about trend, T_t, autocorrelations of nearby past values and autocorrelations of seasonal values, $C_t + N_t$. In step 2, the seasonalities, S_d, of D_t are separated from the trend using STL, this seasonal component is added back after trend interpolation. The idea is the use of the seasonally adjusted of D_t for ARIMA so that it is able to capture autocorrelation of nearby values only. The de-seasonal series D_t' allows avoiding performing seasonal differencing, which has less meaning for such high frequency seasonalities (compare the current demand to a far distant point of time in the past). The seasonal difference, in the context, is the change between the demand of the current hour to the demand of the same hour of the same day of week last week $d = 336$, or to the demand of the same hour of yesterday $d = 48$. Although such comparison makes sense in some low-demand hours of the day, during high-demand hours, the information about the recent consumption past 30-minutes intervals is more reliable to tell something about demand of a specific time. This reaches consensus on finding in case of MCAR, where ARIMA + Kalman provides competitive results. Furthermore, the differencing requires many data points as long periods of seasonal. No values to subtract makes the values at the beginning of series are not available. The interpolation, obtained from STL decomposition, is possible because of the underlying LOESS (Locally Estimated Scatter Smooth). At each point, the response is approximated by fitting a low-degree polynomial to a set of nearest neighbours of that point [21]. The smoothing of response given nearest neighbours as regressors allows the LOESS curve is defined anywhere, and is able to handle the missing, so that the seasonal component of dynamics, S_d, does not have a gap. The seasonally adjusted of D_t, however, has gap, is fitted by ARIMA and is interpolated by Kalman Filter and Smoother. The purpose is to extract the autocorrelation information to do interpolation for fitted series. Actually, fitting univariate local linear trend series in discrete time using a Kalman filter is equivalent to fitting a spline [12] (section 3.9). Spline smoothing approximates the trend of dynamics, T_d. Re-seasonal the interpolated trend, T_d, then add the seasonal component, S_d, back to obtain the imputation. Now, the residual of dynamics, ϵ_t, is minimalized and is nearly white noise, which means the DHR adaptation exploited as much information as possible to impute.

The dataset Calls reveals a questionable nature of trend interpolation for univariate series.

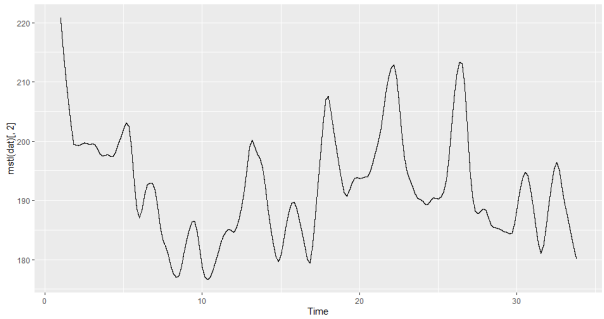

Fig. 13. Trend of Calls

Theoretically, trend imputation for univariate can be any curve as there are no external covariates to interpolate. The trend of Calls obtained from STL decomposition seems not able to interpolate because of the sudden peaks and troughs at different levels. Sometimes, the simple extrapolation of regression with Fourier terms is enough to impute. This reminds that imputation for MAR greatly depends on the gap position. If the gap begins and ends at a part of trend, somewhat 'unpredictable' from past observations where ARIMA from step 2 works well, any methods become unreliable.

The consequence is experiments evaluation may be misleading. The reason is, in simulation, trend is known, however, is unknown in real missing case. A small RMSE (MAE) in simulation, does not mean that continues to be small in a real unknown trend case, simply because there is no "correct" trend to measure. It is no longer a certainty that interpolation for trend performs better than extrapolation, and sophisticated method performs better than simple one.

VI. Conclusion

A. Answers to research questions

Answer to question 1 I-E2: To impute seasonalities, DHR employs Fourier terms of different frequencies. In the suggested method uses STL to, once again, obtain seasonalities of dynamics, and the trend interpolation employs a Kalman filter and smoothing for an ARIMA model. In DHRTVPs, the Fourier terms and Kalman filter and smoothing for State Space form of the series.

Answer to question 2 I-E2: The seasonal component is well captured by both Fourier terms and STL decomposition. The residuals are nearly white noise indicates that the suggested algorithm makes use of information in the series to impute. This result in high degree of accuracy imputation (missing rate less than 10%, MAR & MCAR) and good accuracy rate imputation (missing rate is between 10% and 30%, MCAR) for Taylor dataset. Especially, in MAR, where imputation for the seasonality component matters the most, adaptation of DHRFT outperforms other methods at any rate.

Answer to question 3 I-E2: In MCAR, DHRFT gives high degree of accuracy at missing rate less than or equal to 10%, good accuracy at missing rate between 10% and 30%, reasonable imputation at rate between 30% and 50%. When the rate is higher than 50%, the imputation is inaccurate.

In MAR, the suggested adaption of DHRFT gives high degree of accuracy at missing rate less than 10%, reasonable imputation at missing rate between 10% and 20%, and inaccurate results for rate higher than 30%.

The MAPE interpretations above (for Taylor) is invalid for series that contains many near-zero values, or complicated trend (like Calls).

B. Societal implications and Recommendations

DHR can be used to impute series with long seasonal periods and multiple seasonalities. To select the best method, one needs to pay attention to assumptions: whether seasonal autocorrelations have strong impact (if not, choose a normal ARIMA instead of DHRFT), whether (near-)zero values are in the series (one should revise near-zero imputing values, probably, set some to zero), and the "regular" shape of the trend (it is better to assume a stochastic trend). In MCAR, ARIMA is recommended for its simplicity and good result. In MAR, adaptation of DHR is recommended, however, one should be sceptical when impute a large gap (missing rate $> 10\%$), and should restrain the imputation for series with missing rate $< 30\%$ only. It is recommended to effectively exploit the advantage of DHR by including external variables whenever possible, as domain knowledge plays a crucial role in determine the trend interpolation, and evaluate quality. One should not solely rely on metrics to judge, thoroughly checking for assumptions (and residuals plot) is recommended.

VII. Acknowledgment

I would like to thank my supervisors Dr. Joël M.H. Karel and Dr. Ralf Peeters for all their help and advice with this thesis. I also appreciate all the support I received from my family. Lastly, I would like to thank the IconPro company in Aachen, Germany for the internship that allowed me to conduct this thesis.

VIII. References

References

[1] Engels, J. (2003). Imputation of missing longitudinal data: a comparison of methods. Journal of Clinical Epidemiology, 56(10), 968–976. https://doi.org/10.1016/s0895-4356(03)00170-7

[2] Waal, D. T., Pannekoek, J., & Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation (1st ed.). Wiley.

[3] Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). Otexts.

[4] Rubin, D.B. Inferences and missing data. Biometrika. 1976; 63: 581-590

[5] Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, & Jörg Stork. (2015). Comparison of different Methods for Univariate Time Series Imputation in R. ArXiv: Applications.

[6] Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data (Chapman & Hall/CRC Monographs on Statistics and Applied Probability) (1st ed.). Chapman and Hall/CRC.

[7] 6.4.4.4. Common Approaches to Univariate Time Series. (n.d.). https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc444.htm

[8] Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). Time Series Analysis, Forecasting and Control, 3rd ed. Prentice Hall, Englewood Clifs, NJ.

[9] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 38(18), 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

[10] Walter.O, Y., Kihoro, J., Athiany, K., & KibunjaH., W. (2013). Imputation of incomplete non-stationary seasonal time series data. Mathematical theory and modeling, 3, 142-154.

[11] Peter C. Young, Diego J. Pedregal, & Wlodek Tych. (1999). Dynamic harmonic regression. Journal of Forecasting, 18(6), 369–394. https://doi.org/10.1002/(sici)1099-131x(199911)18:6

[12] Durbin, J., & Koopman, S. J. (2012). Time Series Analysis by State Space Methods (Oxford Statistical Science Series) (2nd Revised ed.). Oxford University Press.

[13] Harvey, A. C. (1993). Time Series Models: 2nd Edition (second edition). The MIT Press.

[14] statsmodels.tsa.forecasting.stl.STLForecast — statsmodels. (n.d.). https://www.statsmodels.org/dev/generated/statsmodels.tsa.forecasting.stl.STLForecast.html

[15] Forecasting using stl objects — forecast.stl. (n.d.-b). https://pkg.robjhyndman.com/forecast/reference/forecast.stl.html

[16] Zavala, A. J., & Messina, A. R. (2014). A Dynamic Harmonic Regression Approach to Power System Modal Identification and Prediction. Electric Power Components and Systems, 42(13), 1474–1483. https://doi.org/10.1080/15325008.2014.934932

[17] Harris, R., & Sollis, R. (2003). Applied Time Series Modelling and Forecasting (1st ed.). Wiley.

[18] Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society, 54(8), 799–805. https://doi.org/10.1057/palgrave.jors.2601589

[19] De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. Journal of the American Statistical Association, 106(496), 1513–1527. https://doi.org/10.1198/jasa.2011.tm09771

[20] McKnight, K. M., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). Missing Data: A Gentle Introduction. Guilford Publications.

[21] R. B. Cleveland. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. Journal of Office Statistics, 6(1), 3–73. http://ci.nii.ac.jp/naid/10014960531

[22] Buuren, S. van, Taylor & Francis Group, & van Buuren, S. (2021). Flexible Imputation of Missing Data, Second Edition. Taylor & Francis.

[23] C. D. Lewis. (1982). Industrial and business forecasting methods: a practical guide to exponential smoothing and curve fitting. Butterworth Scientific EBooks.

[24] Winters, P. R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. Management Science, 6(3), 324–342. http://www.jstor.org/stable/2627346

[25] Harvey, A., & Koopman, S. J. (1993c). Forecasting Hourly Electricity Demand Using Time-Varying Splines. Journal of the American Statistical Association, 88(424), 1228–1236. https://doi.org/10.1080/01621459.1993.10476402
[26] Jones, R. H. (1980). Maximum Likelihood Fitting of ARMA Models to Time Series With Missing Observations. Technometrics, 22(3), 389–395. https://doi.org/10.1080/00401706.1980.10486171
[27] Kohn, R., & Ansley, C. F. (1986). Estimation, Prediction, and Interpolation for ARIMA Models with Missing Data. Journal of the American Statistical Association, 81(395), 751–761. https://doi.org/10.1080/01621459.1986.10478332
[28] Cryer, J. D., & Chan, K. (2008). Time Series Analysis: With Applications in R (Springer Texts in Statistics) (2nd ed.). Springer.
[29] Harvey, A. C., & Peters, S. (1990). Estimation procedures for structural time series models. Journal of Forecasting, 9(2), 89–108. https://doi.org/10.1002/for.3980090203

APPENDIX

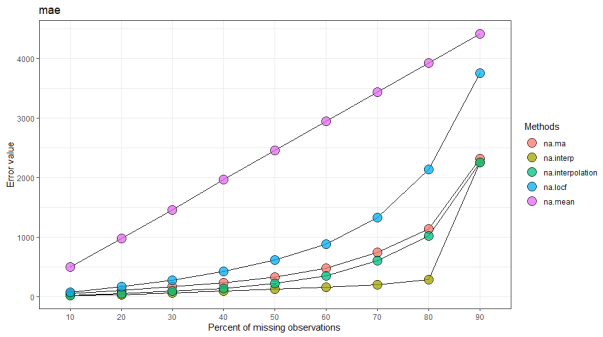

Fig. 14. MAE: Simple imputation methods performance (taylor)

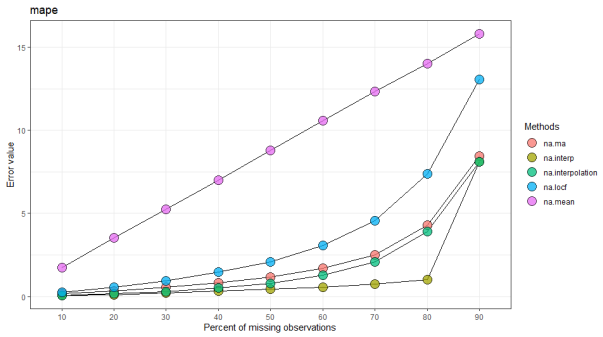

Fig. 15. MAPE: Simple imputation methods performance (taylor)

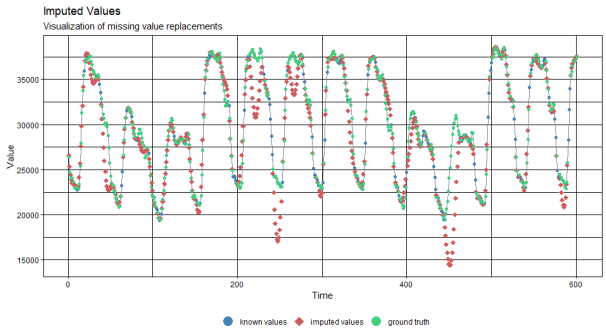

Fig. 16. MCAR, 70%: ARIMA + Kalman imputation (taylor)

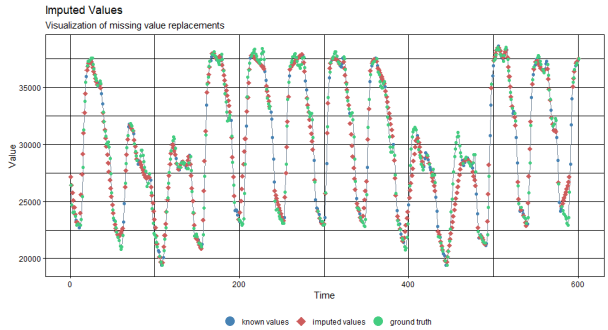

Fig. 17. MCAR, 70%: DHRFT + Kalman imputation (taylor)

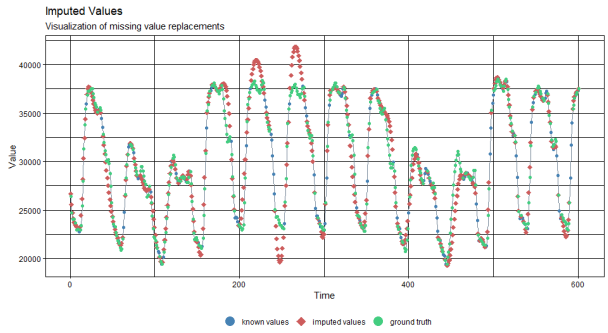

Fig. 18. MCAR, 70%: BSM + Kalman imputation (taylor)

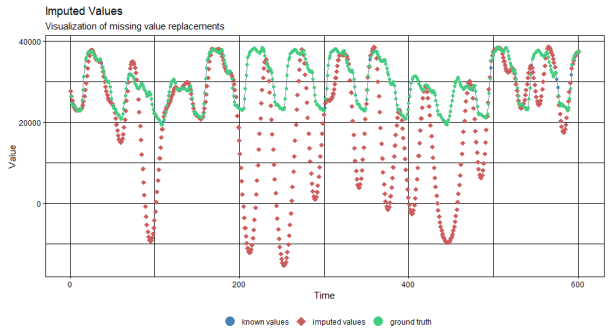

Fig. 19. MCAR, 90%: ARIMA + Kalman imputation (taylor)

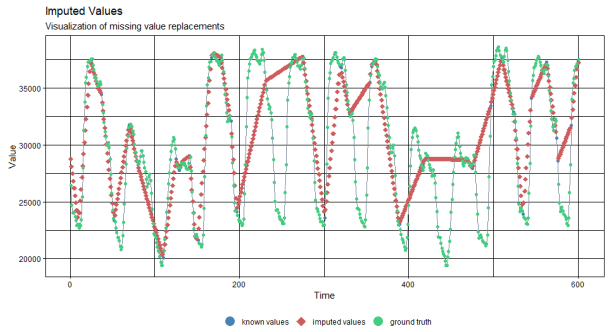

Fig. 20. MCAR, 90%: BSM + Kalman imputation (taylor)

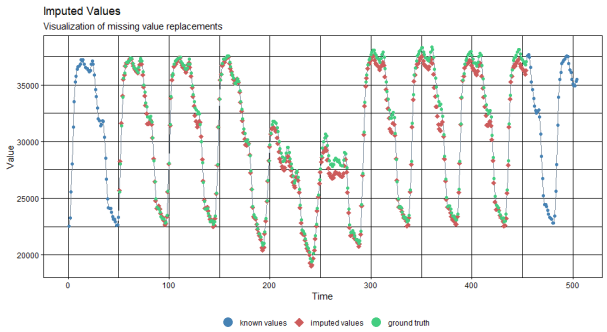

Fig. 21. Best imputation method for MAR, 10%: Regression with dummy seasonal trend (taylor)

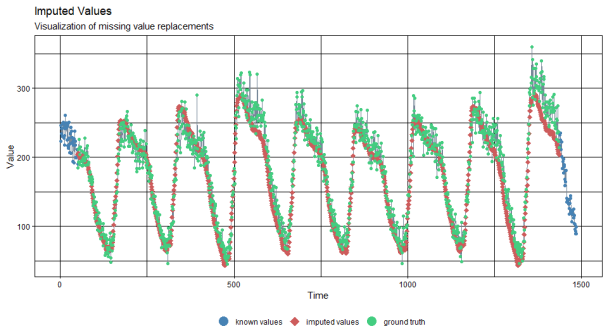

Fig. 22. MAR, 5%: DHRFT (calls)

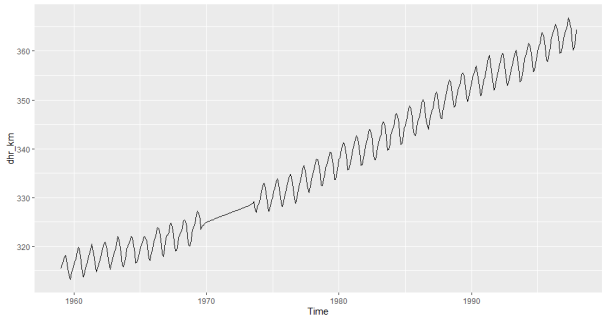

Fig. 23. MAR, 10%: DHRFT + Kalman (CO_2)

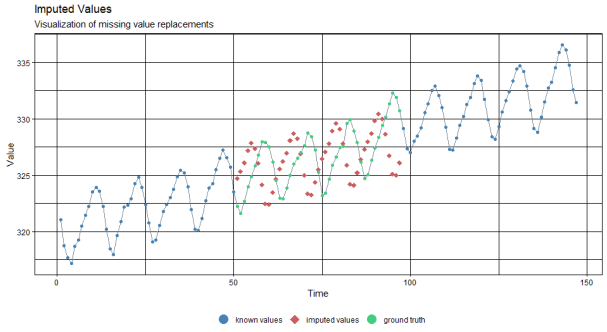

Fig. 24. MAR, 10%: DHRFT (CO_2)

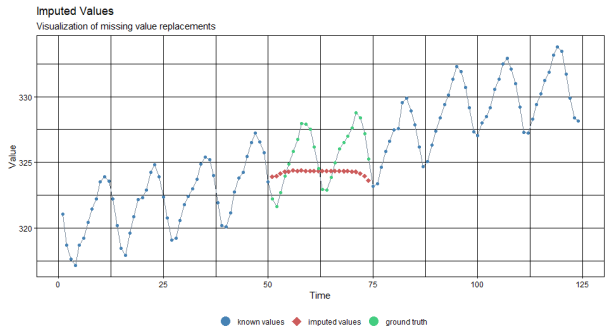

Fig. 25. MAR, 5%: DHR + Kalman (CO_2)