Probabilistic Imputation for High Resolution Univariate Electric Load Data with Large Gaps

Patrick Giles
Northern Energy Innovation
Yukon University
Whitehorse, Canada
pgiles@yukonu.ca

Michael Ross
Northern Energy Innovation
Yukon University
Whitehorse, Canada
mross@yukonu.ca

***Abstract*—The use of electric load data within power engineering applications is critical. Such data often contain missing observations, especially higher resolution datasets important for detailed modeling and simulation. Missing data frequently are handled by replacement with new values, in other words imputation. However, most readily available imputation methods will perform unsatisfactorily on electric load data when many successive observations are missing as they cannot capture the periodic variation. As well many methods only provide single point estimates, allowing for no assessment of probabilistic characteristics of the missing data. In this study a new imputation method is proposed that captures the periodic variation common in high resolution load data, as well as generate probabilistic and point estimates for the missing data. The method is evaluated on three real-world high resolution load datasets and compared with a typical imputation technique.**

***Index Terms*—Electric Load, Imputation, Missing Data**

I. Introduction

Electric load data are important in a variety of power engineering applications, such as forecasting future electricity demand [1], integrating renewable power systems [2] [3], or sizing power system components [4] [5]. In addressing these applications, electric load data will be used to inform models or simulations of varying design and complexity. As such, the problem of missing observations in electric load data can be serious depending on the nature of the missingness and the application the data is required for. For some applications a higher resolution of data may have more utility, however it also will be more likely to be incomplete. Missing data may arise from malfunctions in sensors, disruptions in data-recording software, other technical issues, or rejection of recorded data due to obvious error.

The missingness within electric load data will typically be characterized in two forms: (1) sporadic points where data is missing and (2) periods of successive points of missing data, or "gaps". Sporadic missingness is relatively straightforward to handle; interpolation or smoothing techniques will use the local data surrounding the missing points to impute new data. However, when there are large gaps of missing data (due to a sustained malfunction, interruption of power, equipment failure, etc.) interpolation or smoothing techniques will be inadequate for high resolution load data. This is due to the periodic variation, that is seasonality, inherent in load data; across a large gap of missing data there will be cyclical variation unable to be captured by a simpler method. To capture this behaviour more sophisticated techniques that employ the time domain of the data are needed. However, as noted in [6], many out of box techniques perform well imputing missing data with trend or seasonality, but not both.

Another drawback to many readily available imputation techniques is they only provide point estimates. Point estimates will give no information regarding the certainty of the imputed data, and generally will underestimate the variance of the true data, described in [7] and demonstrated in [8] with a multivariate air quality dataset. Depending on the application for the incomplete data it may be desirable to have bounds on the likely range of the imputations with associated confidence levels. For example, consider a load dataset with a large gap of missing data being used to model an isolated power system for the purpose of designing a new solar plant. Predicting the occurrences of community load varying outside the new plant's operating limits would be well served by knowing a likely upper and lower bound on the imputed data. The process of providing probabilistic estimates for missing data rather than point imputations is known as Multiple Imputation (MI). The basic idea behind MI is to generate many complete datasets from the original incomplete dataset, then pool these complete data into an average estimate with corresponding uncertainty estimates about the average [9, pp. 19-20]. A driving idea behind MI is "imputation is not prediction" [9, Sec. 2.6, pp. 55-57]; instead of committing to a single imputation MI generates a set of plausible estimates and quantifies their certainty.

This study seeks to employ MI techniques to generate probabilistic information about the imputed data. The proposed method exploits the seasonality inherent in electric load data through sub-sequences defined by the dominant seasonality in the data, an approach used in typical seasonal time series models. However, established methods cannot handle high frequency seasonality so easily, and methods that have been developed to do so are not uniformly accessible to all users for imputation purposes. The proposed method seeks to provide a more accessible framework for generating imputations for high resolution seasonal data. This study is restricted to the imputation of high resolution univariate load data, logged at

a one-minute period. Multivariate data are not guaranteed to have sufficient correlation between variables to employ popular imputation methods. Additionally univariate data may be all that are available in certain applications due to cost constraints or other barriers. This study also assumes no historical data are available for similar reasons.

In Section II related imputation techniques are reviewed. In Section III-A and III-B fundamental time series models are reviewed and challenges with high frequency seasonality discussed. The proposed imputation method is detailed in Section III-C. The proposed imputation method is then assessed and discussed in IV with real world load data from three isolated communities. In Section V some conclusions are drawn.

II. Related Work

Imputation techniques that explicitly tackle univariate data with large gaps of missing data are sparse in the literature, particularly with respect to multiple imputation.

In [10] a technique is proposed to handle large gaps in univariate data through exploiting the seasonality inherent in a single variable within a cellular usage dataset to split univariate data into multivariate data along the most relevant cycle; the correlational structure in the new dataset is used to justify feeding the new data into the well known multivariate Multiple Imputation by Chained Equations (MICE) algorithm. This approach was shown to outperform a default MICE implementation with other variables in the datase, as well as a Kalman filter. By employing MICE there is a well understood probabilistic component to the imputed values.

In [11] large gaps of missing data within a large industrial sensor dataset were simulated to study an iterative technique that utilized the well known Seasonal Trend Loess (STL) method. A gap of missing data were segmented, with segments iteratively imputed based on the trend and seasonal components of the STL model; as each segment is imputed the accuracy of the remaining segment imputations should be improved. This technique was shown to outperform a default STL decomposition and imputation, however the imputations had no probabilistic interpretation.

In [6] a method that synthesizes the STL algorithm and a seasonal moving window algorithm was presented, with an explicit focus on large gaps of missing data. Before and after a gap of missing data a window was placed and extended such that non-mising data that had been de-trended (but not de-seasonalized) with STL would be included; the window is shifted throughout the remaining de-trended treated data to find the closest match with past values with respect to the root mean square error. The past data that best matched the window are imputed directly. Once again there is no probabilistic interpretation for the imputed data. Further, this method assumes that the local data about the gap are good predictors of the missing values, as well the imputed values being already recorded data may lead to biased imputations.

In [12] a technique is developed to model time series data with high frequency seasonality, although imputation is not explicitly considered. Imputations could be generated through the use of forecasting and backcasting, although this would increase the time and complexity requirements for users.

With the techniques developed in [11] and [6] there was no probabilistic interpretation. In [12] high frequency seasonality is handled, but without considering imputation. In [10] the probabilistic features come courtesy of using functions included in the MICE package for R. While MICE is a popular technique in some applications, the proposed method seeks to reduce complexity and any barriers to understanding which may be present when using advanced prepackaged techniques such as MICE.

III. Methodology

A. Modeling Time Series

Modeling time series data requires specialized methods compared to standard statistical tools. This is because past values in a time series dataset are natural predictors of future data, violating the independence assumption needed for typical statistical modeling. A time series x_t can be thought of as a function of trend, seasonality, and error. These components may be additive or multiplicative.

$$x_t = \begin{cases} m_t + s_t + \varepsilon_t \\ m_t s_t \varepsilon_t \end{cases} \tag{1}$$

Where m_t, s_t, and ε_t are the trend, seasonal, and error components respectively. Trend is a slowly changing function of t, seasonality is a component that repeats through the data with a fixed period, and error is stationary in the sense that the variance is finite, with mean and autocovariance constant with respect to t. A critical step when modeling time series is to remove deterministic components such that the data which remains can be treated as random error, or some function of random error via a structural time series model. The fundamental structural models are autoregressive (AR) and moving average (MA); with AR modeling dependence is assumed in the realized values from the time series and in MA modeling dependence is assumed in the error component of the time series. The ARMA model considers dependence in both realized values and error components. The order (AR(1), MA(1), ARMA(2,3), etc.) of these models references the amount of time lags the dependence is assumed over. An overview of the notation is given in Table I.

B. Challenges with high resolution seasonal data

When considering the additive form of (1), the seasonal component is assumed to be fixed across cycles. This is frequently a poor assumption to make; in the context of this study's high resolution electric load data the i^{th} minute of

TABLE I: Fundamental Time Series Models

AR(P)	$x_t = \nu + \phi_p x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t$
MA(Q)	$x_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$
ARMA(P,Q)	$x_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \nu + \phi_p x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t$

the j^{th} day likely will not be the same as the $j^{th}+1$ day and so on, although they will be highly correlated. A well established method within the ARIMA framework (the "I" refers to the order of integration of the model, the number of differences taken to remove trend and ensure stationarity) to address this is the SARIMA method ("S" refers to seasonal). SARIMA assumes that the seasonal periods within the data can be modeled as realizations from an ARMA process. Stopping here would go too far from the additive model by assuming the seasonal components are totally uncorrelated, so the detrended non-seasonal components are fitted with another ARMA process [13, Sec. 9.6, pp. 310-316]. However many SARIMA processes, and other seasonal time series models, are designed to accommodate simple seasonality with a small periodicity [12]. High resolution electric load data will have seasonality of a very high period; in the case of minutely data, a daily periodicity of 1440. Existing methods to handle the large periodicity may not be readily available or interpretable for all users. For the problem of imputation specifically, generating suitable data from these methods may be cumbersome, potentially overly so for what would be just an initial step in whatever application is required for the load data.

C. Proposed Imputation Method

To illustrate the proposed algorithm the notation will conform to the dimensions of the data evaluated in this study. The process is easily generalizable to univariate data of differing dimensions. Consider an electric load time series x_t measured per-minute logged over a year, such that $t = \{1, \ldots, 525600\}$. Define the missingness of the data by,

$$\{a, b\} \in t, \quad a < b \text{ and } b - a = G \tag{2}$$

where a and b are the end points of the missing gap, and G is an integer defining the length of said gap. Subset the data x_t by seasonality such that

$$x_t = X_{ij} \quad i = \{1, \ldots, 1440\}\ j = \{1, \ldots, 365\} \tag{3}$$

where i and j refer to the minute and the day respectively. This results in a new variable for every minute of every day. Each new variable only takes a single point from a day, all within-day variation is omitted and between day variation is captured for each minute. However, due to the high resolution $X_{i*}, X_{i+1*}, \ldots$ are typically extremely correlated.

To model the trend across the gap of missing data, a running median is fitted for all X_{i*} giving an estimate of trend $\mathbf{m}_i$. The running median's smoothing parameter k is selected to most effectively capture the overall movements in trend while being wide enough to provide a trend estimate for the missing data. For example, $k = 31$ corresponds to a running median over 31 days. By reversing the subset operations in (2) on the $\mathbf{m}_i$ an estimate of the deterministic component m_t and s_t for the missing data will be obtained, shown in Fig. 1. Generating an estimate of trend piecemeal across the seasonal components

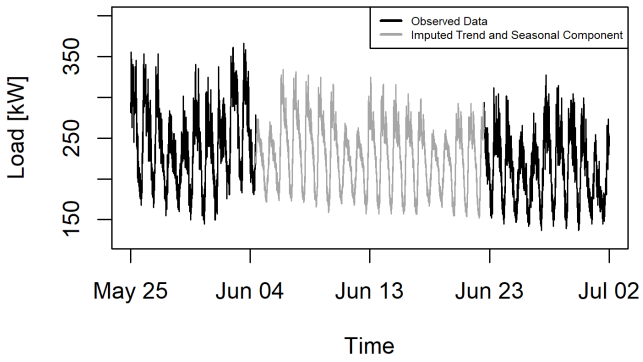


Fig. 1: Trend and seasonality imputation

is a simple way to obtain an estimate of the raw data's overall trend, as well as the seasonality.

Let $\mathbf{e}_i$ represent the residual data after accounting for the trend,

$$\mathbf{e}_i = X_{i*} - \mathbf{m}_i \tag{4}$$

where $E(\mathbf{e}_i) = 0$. Taken individually, the $\mathbf{e}_i$ is assumed to be stationary enough to model as independent and identically distributed noise. However, the correlation between the $\mathbf{e}_i$ has not changed and will still be very large. In other words, the detrended load over the days of the year will be very similar from one minute to the next. To account for this, reverse the subset operations in (2) on $\mathbf{e}_i$ such that

$$\varepsilon_t = \phi\varepsilon_{t-1} + \delta_t \tag{5}$$

where ϕ controls the correlation from one minute to another and δ_t is an error term; note that while ε_t is treated as error within (1), at this stage it is treated as realizations of a data generating process. If $\phi = 1$ then (5) is a random walk. This is undesirable because the correlation is expected to be close to but not exactly unity. While $\phi \neq 1$, (5) is an AR(1) process.

In order to generate imputations the stochastic component of (4), δ_t, is estimated with respect to the minute i, and random draws taken from the underlying probability density. This is achieved through Kernel Density Estimation (KDE). KDE is a well known approach to generate non-parametric probability densities. KDE is used rather than a parametric density like the Gaussian in order to flexibly capture non-typical behaviour in the data, as well as bypass a potentially restrictive parametric form for the stochastic component; what parametric density is a good fit for one dataset may be a poor fit for others. To generate the KDEs, take the difference of successive minutes of detrended data,

$$\mathbf{d}_i = \mathbf{e}_{i+1} - \mathbf{e}_i \tag{6}$$

such that when $i = 1440,\ i + 1 = 1$ and $E(\mathbf{d_i}) = 0$ for all i. Note that through (6) it is assumed $\phi = 1$, however since the parameter of interest from (6) is $\text{Var}(\mathbf{d}_i) = \sigma^2_{\delta_i}$ and it is

already known ϕ will be close to unity it is assumed to not overly bias the result. KDE works through placing a scaled symmetric probability density centered about each $d \in \mathbf{d_i}$. By summing the scaled densities, an estimate of the overall density $\hat{f}_{\delta_i}$ for each $\mathbf{d_i}$ is obtained. Since straightforward KDE will place a probability density over every datapoint, outliers are removed by excluding all $d \in \mathbf{d_i}$ lying outside 5 Median Absolute Deviations (MAD) from the median of $\mathbf{d_i}$; the number of MADs was chosen to comfortably include all the important information, while exclude clear outliers. It is well known that the choice of scaled density in KDE matters little relative to the width, or variability of said density [14, Tab 3.1, p. 43]. A well-established rule of thumb is used to automatically set this parameter [14, eq. (3.31), p. 48].

The solution for ϕ in the AR(1) process has a closed form (7). To estimate ϕ, rearrange (7) into (8),

$$\text{Var}(\varepsilon_t) = \frac{\sigma_\delta^2}{(1-\phi^2)} \tag{7}$$

$$\phi = \sqrt{\frac{\text{Var}(\varepsilon_t) - \sigma_\delta^2}{\text{Var}(\varepsilon_t)}} \tag{8}$$

An estimate for $\text{Var}(\varepsilon_t)$ is obtained through the squared MAD, a robust measure of variance. The estimation of σ_δ^2 is obtained by using (6) and taking the average of the sample variance for all i; the data was already handled for outliers with the density estimation so the non-robust sample variance is fine to use.

All the pieces are in place to generate imputations for the missing data. A summary of the procedure:

1. For electric load data x_t with missingness as defined in (2), subset data into X_{ij} as defined in (3).
2. Obtain estimates of trend $\mathbf{m}_i$ with respect to day j across X_{ij}, then reverse subset operations on trend estimates to obtain an estimate of the deterministic component of the missing data.
3. Take difference of data X_{i*} and $\mathbf{m}_i$ to detrend and generate residuals. Model correlation of residual terms according to an AR(1) process given in (5).
   3.1. Parametrize (5) firstly by estimating error component δ with respect to minute i by taking differences between the residual terms as in (6), then generating estimates of density $\hat{f}_{\delta_i}$.
   3.2. Estimate ϕ through (8).
4. For all $t \in [a, b]$, use parametrized (5) to generate $\varepsilon_{t\in[a,b]}$.
5. Add generated $\varepsilon_{t\in[a,b]}$ to deterministic component.

The above procedure can be extended easily to accommodate additional subsetting of the data. Within this study it was noted that the characteristics of load on weekdays versus weekends may be different enough to be modeled separately. An additional layer of subsetting was used to do this within the above procedure, details are omitted for brevity though tweaks were minor.

For the remainder of the study, the proposed imputation method will be referred as Seasonal Subset Median Imputation (SSMI).

IV. Data and Results

A. Generating the missing data

Complete load data from January to December logged at a per-minute frequency was used from three separate remote northern communities within the Canadian Territories. Missing data are induced in the complete data, then imputations from SSMI are compared to the actual values to assess the imputation method. For all three communities, the amounts of missing data induced are 2.5%, 5%, and 10%, giving a total of 9 scenarios to analyze. These gaps of missing data correspond to approximately 9, 18, and 36 days consecutive days of missing data. For each community, a random draw is taken from $t = \{1, \ldots, 525600\}$ to center the gap where missingness will be induced. If the center of the missing data gap is too close to the extremities of t, then the missingness is extended into the previous or subsequent year and the data reshaped. Doing this is also important so that the running median has enough information about the gap to successfully capture the underlying trend and seasonality. However, this step imparts an additional assumption on the imputation method: the statistical characteristics of the load data do not greatly change from one year to the next. Depending on the considered data this may be a problematic assumption to make. The amount of missing data also will have an effect on the smoothing parameter of the running median, larger gaps of missing data will require a larger smoothing parameter to provide estimates across the gap.

B. Assessing the Imputations

To assess the performance of the imputed data, five metrics are used.

1. Root Mean Squared Error (RMSE): The RMSE is the square root of the average of the squared differences between the imputed and true values, noted as x_{impute} and x_{true} respectively. The RMSE measures the overall precision of the imputed values.

$$\text{RMSE}(x_{\text{impute}}, x_{\text{true}}) = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(x_{\text{impute}g} - x_{\text{true}g})^2} \tag{9}$$

2. Mean Absolute Error (MAE): The MAE is the average of the absolute differences between the imputed and true values. Like RMSE, it is a measure of the overall precision of the imputations, however it is less sensitive to large deviations in the imputations as the differences are not squared.

$$\text{MAE}(x_{\text{impute}}, x_{\text{true}}) = \frac{1}{G}\sum_{g=1}^{G}|x_{\text{impute}g} - x_{\text{true}g}| \tag{10}$$

TABLE II: Imputation Results

Gap Size	Method	Community A					Community B					Community C				
		RMSE	MAE	SIM	CR	AW	RMSE	MAE	SIM	CR	AW	RMSE	MAE	SIM	CR	AW
2.5	SSMI	**21.3**	**16.8**	0.93	91.9	75.9	**13.9**	11.1	0.90	93.1	49.9	**22.1**	**17.6**	0.93	91.3	76.1
	NAD	21.8	17.3	0.93	-	-	14.0	**10.8**	**0.91**	-	-	22.3	17.7	0.93	-	-
5	SSMI	**21.7**	**17.3**	**0.94**	94.1	81.6	**13.7**	**10.7**	0.91	94.3	53.3	26.2	20.5	0.92	89.1	82.0
	NAD	24.9	19.7	0.93	-	-	14.1	11.0	0.91	-	-	**23.0**	**18.0**	**0.93**	-	-
10	SSMI	**24.7**	**19.8**	0.93	93.4	88.9	**15.0**	**11.8**	0.90	93.2	55.9	**35.5**	**28.4**	**0.90**	80.9	90.5
	NAD	26.6	20.7	**0.94**	-	-	16.0	12.3	**0.91**	-	-	42.3	32.3	0.89	-	-

3. Similarity (SIM): The similarity function specifically evaluates the ability of the imputation method to replicate the true data. Results are scaled to $[0, 1]$, with a value of 1 indicating a perfect replication of the actual data.

$$\text{SIM}(x_{\text{impute}}, x_{\text{true}}) = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{1+\frac{x_{\text{impute}g}-x_{\text{true}g}}{max(x_{\text{true}})-min(x_{\text{true}})}} \tag{11}$$

4. Coverage Rate (CR): The CR is the proportion of confidence intervals estimated over the imputed data that contain the true data. The CR should be close to the nominal level of the underlying confidence intervals. In [9, p. 52] it is quoted that a CR greater than the nominal level is a "lesser sin" than a CR lesser than the nominal level, the former indicating a wider dispersion of estimates than the latter; the author recommends a nominal level of 95% for the CR assessment. In contrast to the RMSE, MAE, and SIM which quantify accuracy in varying ways, the CR assesses the overall statistical quality of the imputations. This is critical within the MI framework, recall "imputation is not prediction"; the goal is not to perfectly replicate missing data but to generate a range of plausible possibilities.
5. Average Width (AW): The AW is simply the average width of the confidence intervals estimated in the CR statistic. Ideally the AW should be as small as possible while keeping the CR close to the nominal level.

The number of imputations generated follows the suggestion of [9, p. 60] at 200, chosen since it is desirable for the purposes of this study to approximate the full distribution of the underlying missing data. The relevant percentiles for each imputed data point are calculated from the 200 samples to generate the confidence intervals in the calculation of the CR and AW measures.

The SSMI is further analyzed through comparison with a straightforward and commonly used single imputation approach used to impute electric load data, the Nearest Average Day (NAD) [15, p. 70]. The details may vary from application to application, but for this study an average of the nearest two weekdays and weekend days before and after the missing data gap is taken, and then values in the gap imputed with the calculated averages with respect to whether the value falls on a weekday or a weekend.

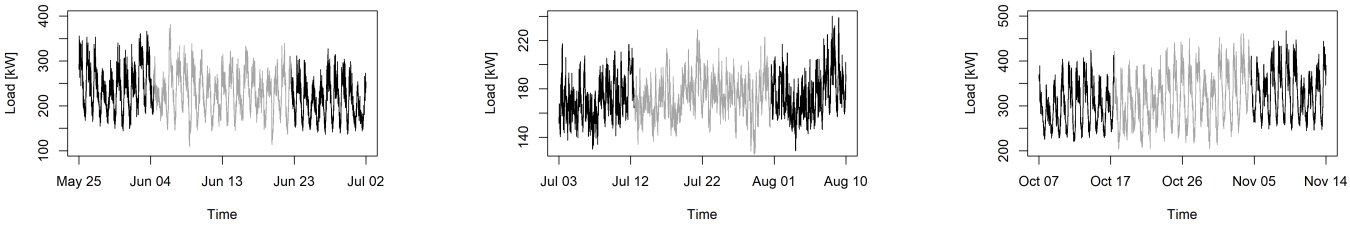
C. Results and Discussion

In Table II the imputations are assessed according to the five metrics, RMSE, MAE, SIM, CR, and AW. All instances where SSMI or NAD outperformed the other the other were bolded. The RMSE and MAE fluctuate steadily with respect to the gap size in communities A and B, with a larger gap of missing data leading to less accurate imputations on an order of 1-3 kW for 5% to 10% gap size, and virtually no change between 2.5% and 5% gap size. In contrast, the accuracy of the imputations markedly declines in community C as gap size increases. As well, the CR and AW perform poorly at 2.5% and 5% gap sizes, and completely deteriorates at 10% gap size. The CR and AW are satisfactory for communities 1 and 2, underperforming the nominal confidence interval coverage of 95% in all cases but not overly so. Interestingly the CR increases from 2.5% to 5% gap size in both community A and B. The SIM performs best overall for community A and worst for community B, but generally remaining consistent between all three communities.

The accuracy of the imputations is acceptable for all three communities despite marked worse performance in C. However, the CR deteriorating at 10% gap size is concerning; further investigation will be needed to determine the cause of the notable statistical underperformance of the imputations in community C. Another point to note is that the smoothing parameter k must increase as the gap size increases to provide estimates of all the missing data. The k were chosen as 21, 31, and 45 for the 2.5%, 5%, and 10% respective gap sizes. A future study will investigate the sensitivity of the selection of the smoothing parameters.

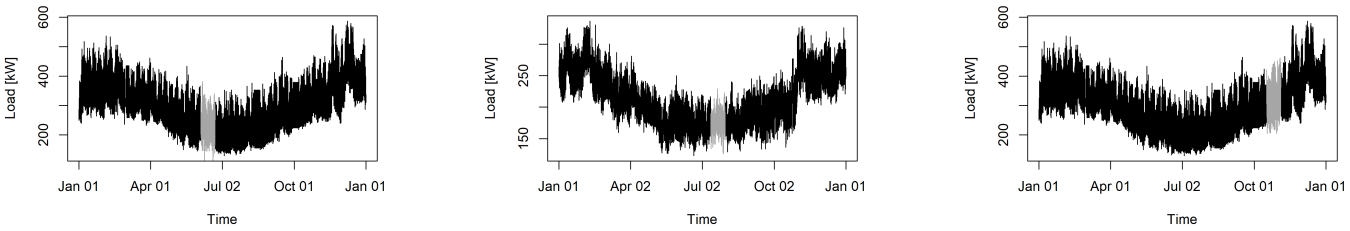
When comparing the SSMI and NAD methods, note that as NAD only produces single imputations, the probabilistic measures are not applicable. Overall the SSMI outperforms NAD, with a notable exception occurring for the 5% gap size in community 3 when NAD outperforms SSMI in all three non-probabilistic metrics. However, given the SSMIs problems in community C as well as that NAD greatly worsens relative to SSMI at 10% gap size, overall the SSMI outperforms the NAD.

To highlight the visual performance of the proposed method, a random imputed series is taken from a set of 200 for all communities at the 5% gap size and plotted with the local neighbouring data in Fig. 2, and the entirety of the data in Fig. 3. The actual data are in black and the imputed data in



(a) Community A (b) Community B (c) Community C

Fig. 2: Snapshot of single imputed series from SSMI method for communities A - C and local surrounding data.



(a) Community A (b) Community B (c) Community C

Fig. 3: Single imputed series from SSMI method for communities A - C and all other data.

grey. The imputed series look satisfactory in Fig. 2, however in Fig. 3c and to a lesser extent in Fig. 3a the lower bound of the imputed data exceeds the local behaviour. While these figures are considering only a single realization of the stochastic component of SSMI, ideally this behaviour would not be present. The greater variability of the imputed data in Fig. 3c may be a clue as to the underperformance of the imputations on a whole for community C in Table II.

V. Conclusions

Within this study a method for generating imputations from high resolution seasonal load data was proposed. The mathematical and statistical detail of the process was outlined step by step. Real world data from three northern communities were used to evaluate the proposed method, with missingness artificially induced in the data to validate the imputations. The results for the proposed method were largely encouraging, with imputations generally following the deterministic and stochastic characteristics of the true missing data. Comparing the proposed method to an established imputation strategy showed satisfactory results, with SSMI overall outperforming the established method's accuracy. However, it is evident further investigation is needed to ensure the method is robust to a greater variety of datasets. The role of the various components of the imputation model on the final results would also be of interest.

In future work, the proposed method will be refined and directly evaluated against more complex imputation methods, as well as forecasts and backcasts from models that explicitly handle high frequency seasonality but are not necessarily tailored to imputation. This comparison will assess not only accuracy, but also the probabilistic characteristics of MI from all considered methods.

Acknowledgment

The authors would like to thank Jason Zrum and Joe Collier for their thoughts and feedback in the development of this work. The authors would also like to thank Northwest Territories Power Corporation, ATCO Electric Yukon, and Yukon Energy Corporation for data used in this study.

References

[1] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy forecasting: A review and outlook," *IEEE Open Access Journal of Power and Energy*, vol. 7, 2020.

[2] M. R. et al., "Grid impact study for old crow solar project," Northern Energy Innovation, Yukon College, Tech. Rep., 2018.

[3] I. Das and C. A. Canizares, "Renewable energy integration in diesel-based microgrids at the canadian arctic," *Proceedings of the IEEE*, vol. 107, 9 2019.

[4] J. Zrum, S. Sumanik, and M. Ross, "Arviat power system impact study," Northern Energy Innovation, Yukon University, Tech. Rep., 2019.

[5] W. Cai, X. Li, A. Maleki, F. Pourfayaz, M. A. Rosen, M. A. Nazari, and D. T. Bui, "Optimal sizing and location based on economic parameters for an off-grid application of a hybrid system with photovoltaic, battery and diesel technology," *Energy*, vol. 201, pp. –, 6 2020.

[6] S. Chandrasekaran, M. Zaefferer, S. Moritz, J. Stork, M. Friese, A. Fischbach, and T. Bartz-Beielstein, "Data preprocessing: A new algorithm for univariate imputation designed specifically for industrial needs," in *Workshop Computational Intelligence*, 2016.

[7] D. B. Rubin, "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, vol. 91, pp. 473–489, 6 1996.

[8] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*, vol. 38, pp. 2895–2907, 6 2004.

[9] S. van Buuren, *Flexible Imputation of Missing Data, Second Edition*. CRC Press, 2018.

[10] A. Chaudhry, W. Li, A. Basri, and F. Patenaude, "A method for improving imputation and prediction accuracy of highly seasonal univariate data with large periods of missingness," *Wireless Communications and Mobile Computing*, vol. 2019, 2019.

[11] Y. Liu, T. Dillon, W. Yu, W. Rahayu, and F. Mostafa, "Missing value imputation for industrial iot sensor data with large gaps," *IEEE Internet of Things Journal*, vol. 7, pp. 6855–6867, 8 2020.

[12] A. M. de Livera, R. J. Hyndman, and R. D. Snyder, "Forecasting time series with complex seasonal patterns using exponential smoothing," *Journal of the American Statistical Association*, vol. 106, pp. 1513–1527, 12 2011.

[13] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. Springer New York, 1991, pp. 310–316.

[14] B. Silverman, *Density Estimation for Statistics and Data Analysis*, 1st ed. Chapman and Hall, 1986.

[15] R. Weron, *Modeling and forecasting electricity loads and prices*, 1st ed. John Wiley & Sons Ltd, 2006.