Transportation Research Part D 38 (2015) 125–143

Contents lists available at ScienceDirect

Transportation Research Part D

journal homepage: www.elsevier.com/locate/trd

Predictive usage mining for life cycle assessment

Jungmok Ma [a], Harrison M. Kim [b,*]

[a] Department of National Defense Science, Korea National Defense University, Seoul, Korea
[b] Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Article history:

Keywords:
Life cycle assessment
Usage modeling
Time series segmentation
Time series analysis

ABSTRACT

The usage modeling in life cycle assessment (LCA) is rarely discussed despite the magnitude of environmental impact from the usage stage. In this paper, the usage modeling technique, predictive usage mining for life cycle assessment (PUMLCA) algorithm, is proposed as an alternative of the conventional constant rate method. By modeling usage patterns as trend, seasonality, and level from a time series of usage information, predictive LCA can be conducted in a real time horizon, which can provide more accurate estimation of environmental impact. Large-scale sensor data of product operation is suggested as a source of data for the proposed method to mine usage patterns and build a usage model for LCA. The PUMLCA algorithm can provide a similar level of prediction accuracy to the constant rate method when data is constant, and the higher prediction accuracy when data has complex patterns. In order to mine important usage patterns more effectively, a new automatic segmentation algorithm is developed based on change point analysis. The PUMLCA algorithm can also handle missing and abnormal values from large-scale sensor data, identify seasonality, and formulate predictive LCA equations for current and new machines. Finally, the LCA of agricultural machinery demonstrates the proposed approach and highlights its benefits and limitations.

Introduction and background

Life cycle assessment (LCA) is an analytical assessment tool to quantify environmental impact of a product or system (Rebitzer et al., 2004; Finnveden et al., 2009). The potential environmental impact can be generated from all the stages of a product, i.e., manufacturing, usage, maintenance, and end-of-life. The LCA approach provides a holistic and systematic way to manage data associated with the product of interest. With the popularity of sustainable design and environmentally conscious design, LCA studies have reported various materials, electronics, automobiles, and complex systems (Kwak, 2012).

The LCA framework (Guinée, 2002; Reap et al., 2008a) consists of goal and scope definition, inventory analysis (LCI, life cycle inventory), impact assessment (LCIA, life cycle impact assessment) and interpretation. The goal and scope definition is the phase that defines the purpose, the systems or products, and the level of sophistication. The LCI is the phase that defines the system boundaries and the flow diagrams with unit processes (e.g., extraction of oil, refining, production of electricity, etc.). The main result from the LCI is the inventory table which quantifies inputs (e.g., raw material, land, energy, etc.) and outputs (e.g., pollutants such as CO_2, SO_2, NO_x, etc.) to the environment. The LCIA is the phase that translates the inventory table into relevant impact categories (e.g., carcinogens, climate change, acidification, etc.) and quantifies the environmental

* Corresponding author at: 104 S. Mathews Ave., Urbana, IL 61801, USA. Tel.: +1 (217) 265 9437; fax: +1 (217) 244 5705.
E-mail address: hmkim@illinois.edu (H.M. Kim).

http://dx.doi.org/10.1016/j.trd.2015.04.022
1361-9209/

impact using weighting and normalization. The interpretation is the phase that evaluates the results from the LCIA and makes recommendations of the LCA study.

Although the LCA approach is mature and has become a widely used method in various industries, it is usually *static* in that time is not considered in the assessment with the implicit assumption of steady-state processes. The necessity of considering time in LCA was discussed in literature. Reap et al. (2008b) provided insightful reviews on the temporal aspects of LCA. Temporal factors such as different rates of emissions over time and seasonal variation of their impacts can influence the accuracy of LCA. Levasseur et al. (2010) showed that the inconsistency in time frames can affect LCA results significantly. Memary et al. (2012) demonstrated that changes of environmental impact over time are useful information for assessing future technology and options. Collet et al. (2014) presented a method to find the most critical flows of information based on dynamic inventory data (i.e., LCI level) and sensitivity analysis. In addition to the aspect of time, spatial variation is another contributor that can significantly affect the accuracy of LCA (Reap et al., 2008b). Local, regional and continental differences can cause different results of LCA.

In this paper, a new perspective of dynamic LCA is proposed to consider time in LCA, especially the modeling of the usage stage. Among the life cycle stages of a product, the manufacturing stage, which is the chosen stage in the majority of LCA studies, can be considered as a one-time event, i.e., time-independent event. Although the dynamic inventory approach (Collet et al., 2014) attempted to relax this (e.g., the impact from material x or process y can be changed over time), the inventory data is considered constant in this study. On the other hand, the usage stage (with maintenance and end-of-life stages) is a time-dependent event, which means the lifespan of a product has a large impact on LCA. Many studies showed that the majority of environmental impact can come from the usage stage over life cycle (e.g., more than 60% for cars (Sullivan and Cobas-Flores, 2001), more than 80% for off-load machinery (product of interest in this paper) (Kwak et al., 2012), and 80–90% for some small electronics (Telenko and Seepersad, 2014)). Therefore, *how to model the usage stage in LCA* is critical and one of the main questions of this work.

Even though the importance of the usage modeling has been recognized among LCA researchers and practitioners, it is rarely discussed in literature. LCA studies in literature usually utilized a constant rate (Lee et al., 2000; Choi et al., 2006; Kwak et al., 2012; Kwak and Kim, 2013; Li et al., 2013) of usage information (hereinafter constant rate method) with the implicit assumption of steady-state processes (e.g., average fuel consumption rate in kg/h, fixed operating hours per month, etc.). This method is simple and easy to apply, but if data has complex patterns (e.g., trend, seasonality and segments), the prediction accuracy of the constant rate method can be significantly reduced. The constant rate method only allows us to calculate life cycle impact in a nominal time horizon, e.g., 10 years as a whole instead from October 2014 to December 2024. This can be an important issue to policy makers and manufacturers when they want to estimate the environmental impact of the future. Fig. 1 shows the expected result from both the proposed model and the constant rate method. Based on the available historical data, a usage (e.g., diesel fuel consumption) model should be built and used for predicting the future usage profile. It can be seen in Section 'Numerical prediction tests for PUMLCA' that the constant rate method can misinterpret the upcoming usage profile while the proposed model is expected to provide higher prediction accuracy with lower variance predictions.

One exception is Telenko and Seepersad (2014) who proposed a usage context modeling technique in LCA using Bayesian network models. The usage context includes human, situational, and product variables. Based on a pre-defined probabilistic network of relevant usage patterns (e.g., weather → usage of electric kettle with probability of x), a usage profile and its variability can be modeled as a form of distribution. However, in order to apply this approach, causal relationships among different usage contexts should be known, which is expressed as a probabilistic network. For example, the usage of agricultural machinery (e.g., crop sprayer, harvester, nutrient applicator, etc.) can be affected by a various usage context (e.g., weather, soil, experience of farmers, price of fuel and crops, machine deterioration). It will be difficult to correlate these variables with specific usage information (e.g., diesel fuel consumption and operating hours). Furthermore, Telenko and Seepersad (2014) did not consider time in LCA.

Alternatively, this study proposes a time series usage modeling technique, predictive usage mining for life cycle assessment (PUMLCA), as shown in Fig. 2. Companies such as Caterpillar (PRODUCT Link™) and John Deere (JDLink™) have

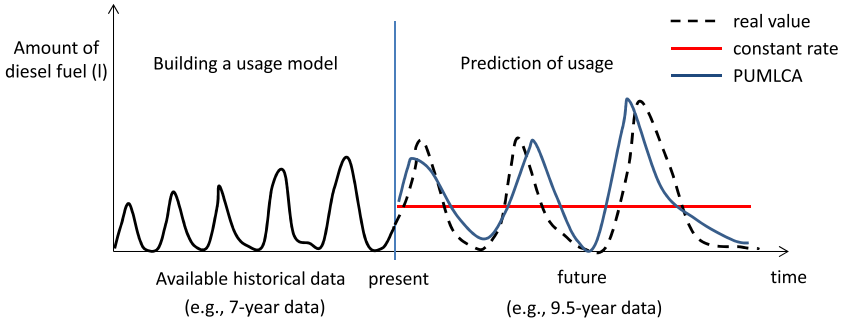

Fig. 1. A prediction scenario of PUMLCA and constant rate method.

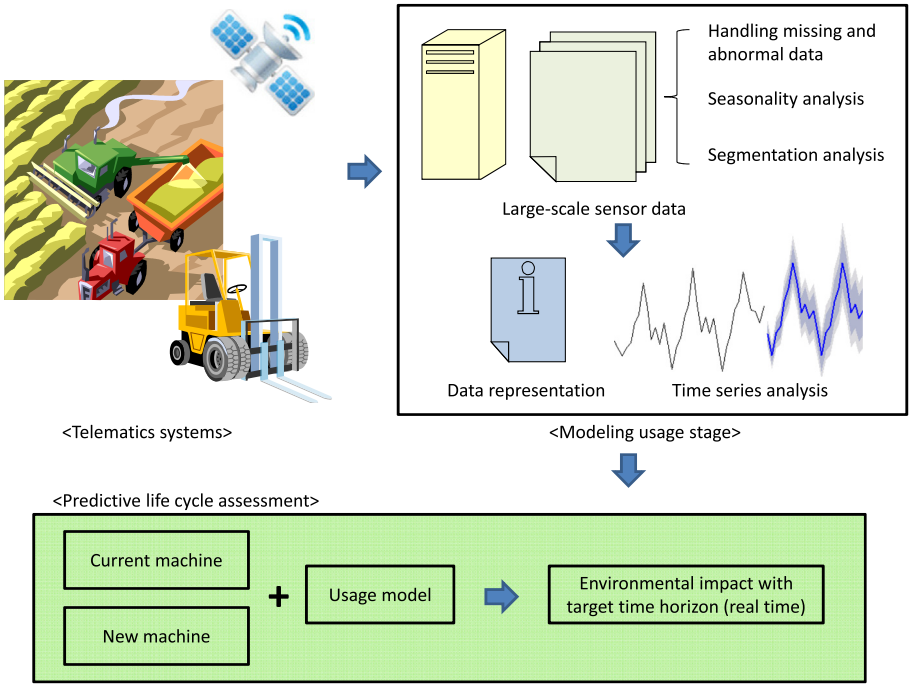

Fig. 2. Overview of PUMLCA.

developed telematics systems for their machinery and have been gathering operational data in real time for various purposes: asset utilization monitoring, location tracking, fleet management, machine health prognostics, etc. These large-scale time-stamped data sets are the sources of data for the PUMLCA algorithm. Usually, the whole picture of a usage profile is not available for currently deploying machines or new machines. Based on the limited past information, future usage patterns should be predicted for LCA as shown in Fig. 1. Time series analysis is useful when future values should be predicted while explanatory variables are difficult to identify. By modeling time series usage information, not only can future usage patterns be obtained, but also variability (i.e., prediction interval). For example, Ma et al. (2014) and Ma and Kim (2014) showed that a trend of valuable information (demand and price) could be mined and reflected in system design using the combination of time series analysis and data mining.

Time series usage information, however, frequently shows highly seasonal activity periods with periodic no-activity or very low-activity periods. For example, combine harvesters are mainly operated during the harvest season with almost zero usage during the off-season. A similar pattern can be observed from seasonally used machinery. This pattern is also widespread for time series data of highly seasonal items such as Christmas, Easter and Halloween products. When analyzing and modeling this kind of time series data, a segmentation can help to find usage patterns more clearly by grouping distinct periods (e.g., off-season period) (Jackson, 2010). Segmentation algorithms (Keogh et al., 2004) were proposed for various

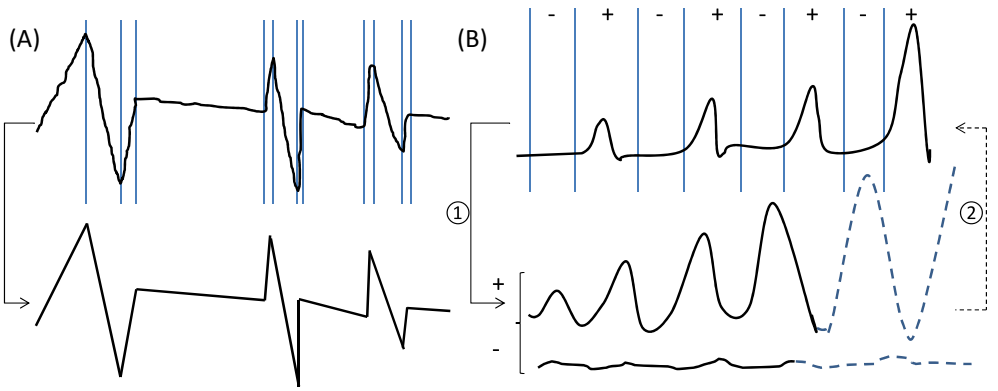

Fig. 3. Time series segmentation (A) piecewise linear representation (redrawn from Keogh et al. (2004)) and (B) segmentation for prediction (redrawn from Hyndman and Athanasopoulos (2013)).

applications such as voice recognition, handwriting recognition, clustering, and classification. However, not much has been reported in the LCA literature whether segmentation algorithms can improve predictive capability. Fig. 3 shows the example. The usual time series segmentation (A (Electrocardiogram) in the figure, piecewise linear representation) is used for the approximation of a time series but the proposed segmentation (B (Monthly sales for a souvenir shop in Queensland, Australia) in the figure, dotted lines for predicted values) is designed to improve the predictive capability of time series modeling by grouping distinct periods and magnifying important patterns (e.g., ① '+' and '−' segments are separated and predicted, ② segments are regrouped with the predicted values). Therefore, *how to segment a time series for better LCA results* is another main question of this work.

The main contribution of this study is to propose the usage modeling technique, predictive usage mining for life cycle assessment (PUMLCA) algorithm, which enables predictive LCA in a real time horizon. The PUMLCA algorithm can provide a similar level of prediction accuracy to the constant rate method when data is constant, and a higher prediction accuracy when data has complex patterns. In order to mine important usage patterns (trend, seasonality and level) effectively from a time series, a new automatic segmentation algorithm is developed based on change point analysis. The PUMLCA algorithm can also handle missing and abnormal values from large-scale sensor data, identify seasonality, and formulate predictive LCA equations for current and new machines. Finally, the LCA of agricultural machinery demonstrates the proposed approach and highlights its benefits and limitations.

The rest of the paper is organized as follows: Section 'Description of predictive usage mining for life cycle assessment algorithm' describes the PUMLCA algorithm. Section 'Design problems with PUMLCA' provides design problems for current and new machines. Numerical prediction tests are presented for PUMLCA and the constant rate method in Section 'Numer ical prediction tests for PUMLCA' followed by a case study of agricultural machinery in Section 'Case study: agricultural machinery'. The benefits and limitations of the proposed methodology along with future research directions are discussed in Section 'Closing remarks and future work'.

Description of predictive usage mining for life cycle assessment algorithm

Fig. 4 outlines the predictive usage mining for life cycle assessment (PUMLCA) algorithm. There are five stages: data preprocessing for handling missing and abnormal values, seasonal period analysis, segmentation analysis, time series analysis, and predictive LCA. Details are explained in each subsection respectively. The algorithm starts from gathering time-stamped sensor data sets with usage information of interest. The amount of fuel (or energy) consumption and operating hours by work modes (e.g., idling and non-idling) are selected as the usage information. In this paper, the usage information is viewed as a result of interactions among human, situational and product variables which are the components of the usage context (Telenko and Seepersad, 2014). For example, the amount of fuel consumed by work modes can be affected by user experience and preference (human variables), weather and soil (situational variables), and machine deterioration and efficiency (product variables). The patterns of the usage information (usage patterns) are defined as trend, seasonality and level in historical time series data. A trend is a long-term increase or decrease pattern; a seasonality is a repeated pattern with a fixed and known period; and a level is base values after removing trend and seasonality. Since a level can be considered as an initial value with a series of random errors, trend and seasonality are the two main patterns that will be mined.

Data preprocessing

After collecting a time series of usage information of interest, it should be checked whether there are missing or abnormal values. Though it is assumed that the error rate of sensor data is very low and the incompleteness of data happens at random, it is still possible to have missing or abnormal values. In order to handle missing values (usually indicated as not available), various imputation techniques are available: (1) removing the missing values, (2) replacing the missing values with random values, adjacent values, mean or median, and (3) replacing the missing values based on values of a correlated variable. Since the volume of collected data is very large, any aforementioned method can be applied.

Unlike missing values, abnormal values (or outliers) are difficult to define. However, similar to the case of missing values, it is assumed that the sample size of abnormal values is much smaller than the volume of the original data and abnormal values are not generated systematically. There are two approaches: (1) three-sigma rule and (2) boxplot. The three-sigma rule states that approximately 99.73% of values lie within three standard deviations of the mean in Gaussian distribution. In other words, if the collected values (y_t) are considered random variables following the Gaussian distribution, abnormal values can be defined as values located outside of Eq. (1):

$$\mu - 3\sigma \leqslant y_t \leqslant \mu + 3\sigma \tag{1}$$

where μ is the mean and σ is the standard deviation.

Another method to detect abnormal values is a boxplot. Abnormal values are defined as values located outside of Eq. (2):

$$Q_1 - 1.5\mathrm{IQR} \leqslant y_t \leqslant Q_3 + 1.5\mathrm{IQR} \tag{2}$$

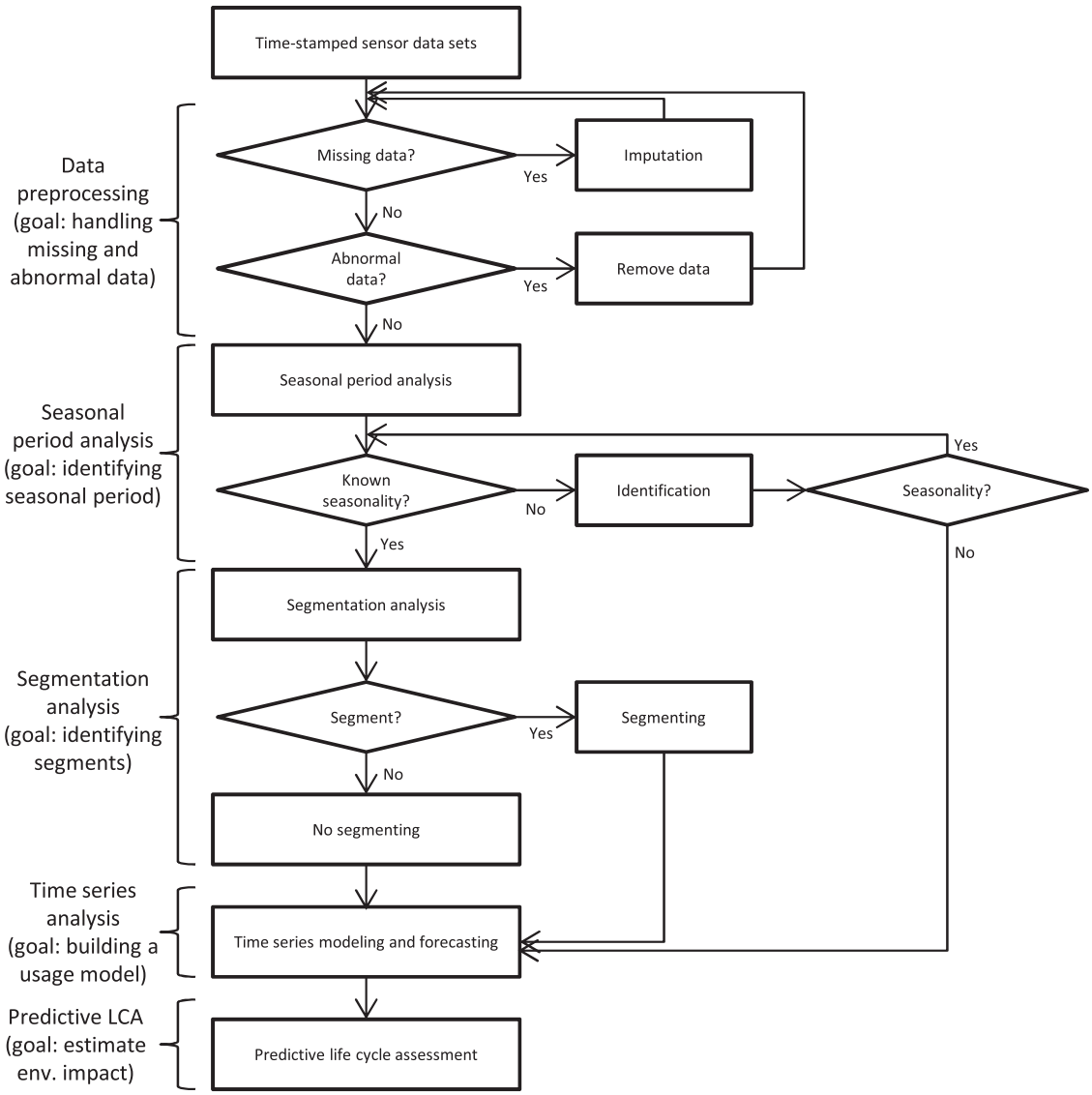

Fig. 4. Overall framework of PUMLCA.

where Q_1 is the 25th percentile, Q_2 is the median or 50th percentile, Q_3 is the 75th percentile, and IQR refers to the interquartile range $(Q_3 - Q_1)$. If data is distributed as the Gaussian distribution, Eq. (2) can be expressed as $\mu \pm 2.7\sigma$. Fig. 4 indicates that detected abnormal values are removed and handled by techniques for missing values.

Seasonal period analysis

The next step is to determine whether there are seasonal patterns, and if there is, what the length (period) of seasonality is. It should be noted that seasonality modeling will be conducted in Section 'Time series analysis', but without the information of the seasonal period, seasonality cannot be modeled. Examples of typical periods include 24 for an hourly series, 7 for a weekly series, 12 for a monthly series, 4 for a quarterly series, etc. If a seasonal period is known, the information can be used. If it is not known, then a dominant period should be identified with different seasonal representations of the original sensor data.

A periodogram (Shumway and Stoffer, 2011) is suggested to identify the important seasonal period. The periodogram is a plot with the *x*-axis of frequencies and the *y*-axis of periodogram values. The periodogram value is a sample spectral density, which can give the relative importance of frequencies. The mathematical expression of the periodogram values are defined as (Shumway and Stoffer, 2011):

$$P\left(\frac{j}{n}\right) = \left[\frac{2}{n}\sum_{t=1}^{n} y_t \cos\left(2\pi\left(\frac{j}{n}\right)t\right)\right]^2 + \left[\frac{2}{n}\sum_{t=1}^{n} y_t \sin\left(2\pi\left(\frac{j}{n}\right)t\right)\right]^2 \tag{3}$$

where y_t is a time series with n discrete time points and (j/n) are frequencies (j cycles in n time points) for $j = 1, 2, \ldots, n/2$. The dominant period (i.e., reciprocal of a frequency (j/n)) can be identified by $\arg\max P(j/n)$.

One helpful treatment before plotting a periodogram is detrending time series usage information (i.e., remove a trend). Two possible methods of detrending will be presented in Section 'Time series analysis'. Also, from a practical standpoint, users can limit a range of frequencies as a meaningful range by their definition.

Segmentation analysis

There are two types of segmentation analysis: deterministic and automatic. Deterministic segmentation analysis can be used when some segments of given time series data show deterministic patterns, e.g., zero usages over time within specific periods. If this prior knowledge is not available or patterns are not deterministic with variable periods, automatic segmentation analysis should be applied. In this paper, a new automatic segmentation algorithm with the change point analysis is presented.

Fig. 5 shows the schematic of the automation segmentation algorithm. A period (n/j) is identified from Section 'Seasonal period analysis' and the number of data points n is proportional to the period (i.e., $n/j = jp/j = p$). For each period, there are p time indexes, $m_1, m_2, \ldots, m_p$. For example, a period 12 has 12 time indexes which are January, February, ..., December. The goal of this algorithm is to find a shared segment (SS) over periods. sp_j denotes a segment, which is a set of p time indexes in the period p_j. The segment does not contain any change point.

Change point analysis is a statistical technique that can detect multiple change points within a time series (Killick et al., 2012). When a discrete time series, $y_{1:n} = \{y_1, \ldots, y_n\}$, is given, positions of change points, $\tau_{1:m}$ ($\tau_0 = 0$ and $\tau_{m+1} = n$) can be identified if the statistical properties of $y_{1:\tau_1}, y_{\tau_1+1:\tau_2}, \ldots, y_{\tau_m+1:n}$ are different in some sense. In this paper, changes in mean are adopted, although changes in variance are another option. In order to identify change points, an objective function is given by Killick et al. (2012):

$$F(n) = \min_{\tau}\left\{\sum_{i=1}^{m+1}[C(y_{(\tau_{i-1}+1):\tau_i}) + \beta]\right\} \tag{4}$$

where C is a cost function for a segment and β is a penalty. For $t < n$, a recursive expression can be determined as follows (Killick et al., 2012) and solved in turn by dynamic programming:

$$F(n) = \min_{t}\left\{\min_{\tau\in\tau_{1:t}}\sum_{i=1}^{m}[C(y_{(\tau_{i-1}+1):\tau_i} + \beta)] + C(y_{(t+1):n}) + \beta\right\} = \min_{t}\{F(t) + C(y_{(t+1):n}) + \beta\} \tag{5}$$

A pruned exact linear time (PELT) method (Killick et al., 2012) was proposed to solve Eq. (5) more efficiently with a pruning procedure instead of searching all possible change points. During iterations for $t < s < n$, only a set of t satisfying Eq. (6) will be considered:

$$F(t) + C(y_{(t+1):s}) + K \leqslant F(s) \tag{6}$$

where K is a constant.

As a cost function, the negative of maximum log-likelihood is used, which is given by Killick et al. (2012):

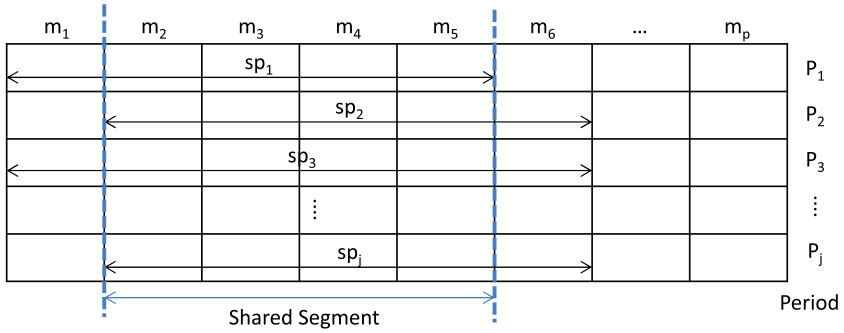

Fig. 5. A schematic of automatic segmentation algorithm.

$$C(y_{(t+1):n}) = -\max_{\theta} \sum_{i=t+1}^{n} \log f(y_i|\theta) \tag{7}$$

where $f(y_i|\theta)$ is a density function with the parameter θ for a segment.

As a penalty, there are some options such as Akaike's Information Criterion (AIC, $\beta = 2p$) and Bayesian Information Criterion (BIC, $\beta = p\log(n)$), where p is the number of added parameters for a change point. It is also possible to specify a type I error (e.g., 0.05 or 0.01) as a penalty value using an asymptotic distribution (Killick and Eckley, 2011). The PELT algorithm is implemented in R (Killick and Eckley, 2011).

The automatic segmentation algorithm based on the change point analysis (i.e., PELT) is provided in Algorithm 1. The goal of this algorithm is to find shared segments over seasonal periods which contain no change point. Unlike the PELT algorithm, change points will be identified within a seasonal period. A penalty, β, should be selected by users. As the penalty value increases, less change points will be identified and the algorithm will be less sensitive over close values. A segment is defined as a group of members within a seasonal period. At least two members are required to be a segment (e.g., $y_{1:3}$ in line 12). In line 4, τ^* contains the possible positions of change points, which are p time indexes within each period (e is the indexes of periods). In lines 5–8 (Killick et al., 2012), the PELT algorithm is implemented with the pruning procedure in line 8. R_{τ^*} is the set of τ^*; τ' is the identified optimal position of change points; CP_e denotes the optimal positions of change points (τ^*) for each period, which is the result of the first part of the algorithm in line 10. Line 12 makes a set of segments, S_e, for each period based on the identified optimal change points (CP_e). Note that $\tau_{1:m_p} = (\tau_1, \ldots, \tau_{m_p})$. Line 13 finds shared segments (SS) over different periods. At this point, it is possible that change points can exist among the sets, S_e, in the shared segments, which indicates that those segments are not similar patterns that repeat periodically. Line 14 makes one new time series (NS) using shared segments of each period (e.g., SS_{p_1} represents the shared segment of the first period). Line 15 applies the PELT method for the new series with no period and a new change point set, CP′, is returned in line 16. The output depends on the new change point set. If there is no change point, the shared segments and the remaining data are grouped as different time series. If there is a change point, no segmentation will be implemented.

Algorithm 1. Automatic segmentation

1: A time series, $y_{1:n}$ with n number of data points
2: A seasonal period, p, where $p = n/j$ with j cycles
3: A measure of fit $C(\cdot)$ and a penalty β
4: **for** $\tau^* = 1, \ldots, m_p$ and $e = p_1, p_2, \ldots, p_j$ **do**
5: Calculate $F(\tau^*) = \min_{\tau \in R_{\tau^*}} \left[\{F(\tau) + C(y_{(\tau+1):\tau^*}) + \beta \} \right.$
6: Let $\tau' = \arg\min_{\tau \in R_{\tau^*}} \left[\{F(\tau) + C(y_{(\tau+1):\tau^*}) + \beta \} \right.$
7: Set $CP_e(\tau^*) = [cp(\tau'), \tau']$
8: Set $R_{\tau^*+1} = \{\tau \in R_{\tau^*} \cup \{\tau^*\} : F(\tau) + C(y_{(\tau+1):\tau^*} + K \leqslant F(\tau^*))\}$
9: **end for**
10: **return** $CP_{p_1}, CP_{p_2}, \ldots, CP_{p_j}$
11: **for** $e = p_1, p_2, \ldots, p_j$ **do**
12: Set $S_e = \{y_{1:\tau_1^*}, y_{(\tau_1^*+1):\tau_2^*}, \ldots, y_{(\tau_{m_p-1}^*+1):\tau_{m_p}^*}\}$
13: Find $SS = \{S_{p_1} \cap S_{p_2} \cap \cdots \cap S_{p_j}\}$
14: Let $NS = \{SS_{p_1}, SS_{p_2}, \ldots, SS_{p_j}\}$
15: Apply line 4 ∼ 9 to NS
16: Get CP′(τ^*)
17: **end for**
18: **return**
19: **if** CP′(τ^*) = null **then**
20: group SS as one time series and remaining as another time series
21: number of time series (s) = z
22: **else**
23: no segmentation, $s = 1$ (i.e., original data)
24: **end if**

Based on the result of the automatic segmentation algorithm, time series analysis methods in the next section will be applied to each segmented time series. Now, each time series has a new period, which is the number of seasonal time indexes.

Time series analysis

Time series analysis includes modeling time series data by extracting important patterns and forecasting future values from the fitted model. The two most widely used time series analysis techniques (Hyndman and Athanasopoulos, 2013)

are adopted in this paper: exponential smoothing (ETS) and autoregressive integrated moving average (ARIMA). Since "each has its strengths and weaknesses" (Hyndman and Khandakar, 2008), either method can be selected by users. Observations are denoted by y_t and a forecast of h ahead time based on all the data up to time t is denoted by $\hat{y}_{t+h|t}$ where h is a real time horizon.

Exponential smoothing

The ETS models refer to an exponential smoothing family (e.g., simple exponential smoothing, Holt's linear trend model, Holt-Winters seasonal model, etc.) based on the innovations state space framework (Hyndman et al., 2008). The ETS model identifies key components of a time series (trend and seasonality) and expresses their relationships (additive and multiplicative) using exponential smoothing.

The simplest model of ETS is given as:

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t) \tag{8}$$

where α is a parameter between zero and one. Eq. (8) represents that the new forecast is the combination of the old forecast and the error from the last forecast. Similar to Eq. (8), there are 30 ETS models with a combination of trend (none, additive, additive damped, multiplicative and multiplicative damped), seasonality (none, additive and multiplicative) and error (additive and multiplicative) (Hyndman et al., 2008).

All the 30 ETS models can be expressed as innovations state space models and the general model is given as (Hyndman et al., 2008):

$$y_t = w(x_{t-1}) + r(x_{t-1})\epsilon_t \tag{9}$$
$$x_t = f(x_{t-1}) + g(x_{t-1})\epsilon_t \tag{10}$$

where x_t is the state vector which contains unobserved components such as the level, trend, and seasonality of a time series; $w()$ and $r()$ are scalar functions; $f()$ and $g()$ are the vector functions; and ϵ_t is the white noise process with variance σ^2. The white noise process is a process that has zero mean, constant and finite variance, and uncorrelated series. Using this innovations state space framework, Hyndman et al. (2008) showed that prediction interval can be obtained along with a point forecast.

In order to get a forecast, $\hat{y}_{t+h|t}$, a recursive expression was summarized as follows (Hyndman et al., 2008):

$$\hat{y}_{t|t-1} = w(x_{t-1}) \tag{11}$$
$$\epsilon_t = (y_t - \hat{y}_{t|t-1})/r(x_{t-1}) \tag{12}$$
$$x_t = f(x_{t-1}) + g(x_{t-1})\epsilon_t \tag{13}$$

Then, a simulation approach (Hyndman and Khandakar, 2008) can be used to simulate ϵ_t for a forecast with a prediction interval.

The remaining part is the identification of trend and seasonality, which is called as the decomposition of a time series. First, the trend component can be estimated ($\hat{T}_t$) by a moving average smoothing. The moving average smoothing of order m is given by Hyndman and Athanasopoulos (2013):

$$\hat{T}_t = \frac{1}{m}\sum_{j=-k}^{k} y_{t+j} \tag{14}$$

where $m = 2k + 1$. The order of the moving average smoothing is a seasonal period, and if the seasonal period is not known, usually odd orders (e.g., 3, 5, 7, 9, etc.) can be applied (Hyndman and Athanasopoulos, 2013). A larger order gives a smoother fit. Then, detrended time series data can be obtained as $y_t - \hat{T}_t$ for the additive model or $y_t/\hat{T}_t$ for the multiplicative model. It should be noted that this is one method to obtain a detrended series for the seasonal period analysis in Section 'Seasonal period analysis'.

Second, the seasonal component can be estimated from detrended series data. An average of each seasonal time index over seasonal periods (e.g., all values in January for monthly data) gives the seasonal component, $\hat{S}_t$.

ARIMA

While the ETS model represents a time series as exponential smoothing of trend and seasonality, the ARIMA model is based on autocorrelations in the time series. The ARIMA model (without seasonality) is a combination of three models given as (Hyndman and Athanasopoulos, 2013):

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta B + \cdots + \theta_q B^q)e_t \tag{15}$$

where the first parenthesis is an autoregressive (AR) model of order p, the second parenthesis is an integration (or differencing operation), and the third parenthesis on the right-hand side is a moving average (MA) model of order q. B represents a backward shift operator, e.g., $By_t = y_{t-1}$.

The AR model of order p is given by:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t \tag{16}$$

where c is a constant and e_t is white noise. This is a linear combination of past observations.

The differencing operation of order 1 and order 2 is given as:

$$y'_t = y_t - y_{t-1} \tag{17}$$
$$y''_t = y'_t - y'_{t-1} \tag{18}$$

The determination of differencing can be made by statistical inference called unit root tests (Hyndman and Athanasopoulos, 2013). It should be noted that this is another method for detrending time series data for the seasonal period analysis in Section 'Seasonal period analysis'.

The MA model of order q is given as:

$$y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \tag{19}$$

This is a linear combination of past forecast errors.

Finally, seasonal ARIMA model can be written as (Hyndman and Athanasopoulos, 2013):

$$\begin{aligned}&(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - \Phi_1 B^m - \cdots - \Phi_P B^{Pm})(1 - B)^d (1 - B^m)^D y_t \\ &\quad = c + (1 + \theta_1 B + \cdots + \theta_q B^q)(1 + \Theta_1 B^m + \cdots + \Theta_Q B^{Qm}) e_t\end{aligned} \tag{20}$$

where lower-case letters p, d, and q are orders for non-seasonal AR, integration, and MA models; upper-case letters P, D, and Q are orders for seasonal AR, integration, and MA models; and m is a period.

In order to forecast future values based on a fitted ARIMA model, Eq. (20) can be expanded so that only y_t will be shown on the left-hand side. By rewriting it as $\hat{y}_{t+h|t}$, a recursive expression can be solved for a forecast of h ahead time.

Close observation for both ETS and ARIMA models reveals similarities. The ETS model starts identifying trend and seasonality and the ARIMA model uses the differencing operation to remove trend and seasonality (i.e., stationarity). The ETS then expresses a series using past level, trend and seasonality with exponentially decreasing weights while the ARIMA expresses a series using past observations and forecast errors.

Automatic modeling of ETS and ARIMA

As shown previously, the ETS and ARIMA require parameter estimation and model selection. Hyndman and Khandakar (2008) provided an automatic forecasting algorithm to handle a large number of univariate time series data. The algorithm is implemented in R package *forecast*. This section briefly introduces the automatic forecasting algorithm for the ETS and ARIMA models.

The automatic forecasting algorithm for the ETS models can be summarized as follows: (1) apply all 30 models and optimize parameters of each model, (2) select the best model based on a penalized likelihood such as AIC and BIC, and (3) forecast future values and obtain prediction intervals based on the selected model.

The automatic forecasting algorithm for the ARIMA can be summarized as follows: (1) apply four possible models and select the best model based on a penalized likelihood, (2) apply 13 variations on the current model and repeat the process if a better model can be identified based on a penalized likelihood, and (3) forecast future values and obtain prediction intervals based on the selected model. Details of these algorithm can be found in the work of Hyndman and Khandakar (2008).

Predictive life cycle assessment

The difference between predictive LCA and original LCA is to model the usage stage (with maintenance and end-of-life stages) as a time series and to forecast future impact in a real time horizon. The total life cycle impact of a product can be expressed as (Kwak and Kim, 2013):

$$I^{total} = I^{mfg} + I^{usage} + I^{maint} + I^{eol} \tag{21}$$

where $I^{mfg}, I^{usage}, I^{maint}$ and I^{eol} represent the impact of manufacturing, usage, maintenance, and end-of-life stage. In the equation, a constant fuel (or energy) consumption rate in the usage stage and replacement cycles in the maintenance stage are components that are dependent upon the expected lifespan. However, the time in Eq. (21) is nominal, e.g., 10 years instead of specifying a time horizon such as from October 2014 to December 2024.

Instead, Eq. (22) gives the total environmental impact in a real time horizon:

$$\sum_{t=i}^{l} I^{total} = I^{mfg} + \sum_{t=i}^{l} \left[I_i^{usage} + I_i^{maint} + I^{eol} \right] \tag{22}$$

where l is the expected life time starting from time i. The impact of manufacturing can be considered as a one-time event while the impacts of usage, maintenance, and end-of-life are affected by time series usage information.

The impact of manufacturing is given as (Kwak and Kim, 2013):

$$I^{mfg} = \sum_r e_r^{raw} N_r + \sum_p e_p^{process} N_p + \sum_s e_s^{trans} N_s \tag{23}$$

where $e_r^{raw}, e_p^{process}$, and e_s^{trans} represent unit environmental impact of raw materials (r), manufacturing processes (p), and transportation (s); N_r, N_p, and N_s denote the number of units of raw materials, manufacturing processes, and transportation.

The impacts of usage, maintenance, and end-of-life are given as:

$$\sum_{t=i}^{l} I^{usage} = \sum_{t=i}^{l} I_i^{fuel} + \sum_{t=i}^{l} I^{emission} = \sum_{t=i}^{l} e^{fuel} N_{ft} + \sum_q \sum_{t=i}^{l} e_q^{emission} ER_q OH_t \tag{24}$$

$$\sum_{t=i}^{l} I^{maint} = \sum_m \sum_{t=i}^{l} e_m^{maint} N_m \left\lceil \frac{\max(OH_t - RC_m, 0)}{RC_m} \right\rceil \tag{25}$$

$$\sum_{t=i}^{l} I^{eol} = e_{used}^{eol} + \sum_m \sum_{t=i}^{l} e_{replace}^{eol} N_m \left\lceil \frac{\max(OH_t - RC_m, 0)}{RC_m} \right\rceil \tag{26}$$

where I^{fuel} and $I^{emission}$ are the impacts of fuel production as in Eq. (23) and emissions while running an equipment; $e^{fuel}, e_q^{emission}, e_m^{maint}, e_{used}^{eol}$, and $e_{replace}^{eol}$ are the unit impacts of fuel, emissions, manufacturing of maintenance part m as in Eq. (23), and end-of-life processing of a used product and replaced part (m); N_{ft} is the amount of fuel consumed per liter; N_m denotes the number of units of part m (in a product); ER_q is the emission rate of emissions q in g/h; OH_t is the operating time in hours; RC_m is the replacement cycle of part m in hours; $\lceil \cdot \rceil$ is the ceiling function. The value of a ceiling function will give the number of replacements for part m. All the unit impacts can be obtained from the ecoinvent database (version 2.2), which is available in the LCA software SimaPro. Note that this study only considers energy-related impacts (e.g., fuel and electricity) of the usage stage, which is identified as the main contributors in literature. Other consumables are not considered, e.g., coffee and water for coffee machines, paper and ink for printers, etc. based on the scope of this study.

Section 'Description of predictive usage mining for life cycle assessment algorithm' described the proposed algorithm from data preprocessing to predictive LCA formulation. Note that the algorithm starts from the available time-stamped data sets (top of Fig. 4) and it is not discussed how many data sets should be available for the algorithm. Empirical studies show that if the available data is not enough to identify useful patterns (e.g., only a few data points), then the result from Secti on 'Automatic modeling of ETS and ARIMA' is identical with the constant rate method, which is smoothing by averaging available data points. Actually, the constant rate method can be considered as a special case of the proposed time series analysis methods. In the next section, the proposed LCA formulation will be elaborated with design problems.

Design problems with PUMLCA

Two system design cases are considered in this study, which is shown in Fig. 6. The first case, analysis for sustainability, is when current machines need to be analyzed for sustainability. In this case, enough usage data is available with manufacturing, maintenance and end-of-life data. Life cycle information includes all the information from life cycle stages and the expected lifespan or target time horizon.

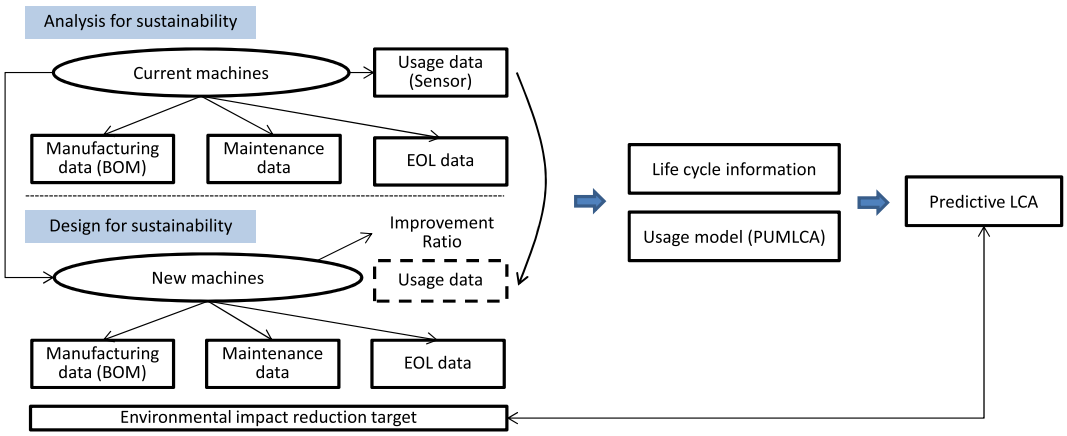

Fig. 6. Two system design cases for predictive LCA.

The amount of fuel consumed, N_{ft}, and operating hour, OH_t, are the time series usage information. The fitted models for N_{ft} and OH_t from ARIMA or ETS are $TS_{ts}^{N_f}$ and TS_{ts}^{OH} with the number of segments s in Algorithm 1. For example, $TS_{ts}^{N_f}$ can be either Eqs. (27) and (28), or Eq. (29):

$$N_{fts} = w(x_{t-1}) + r(x_{t-1})\epsilon_t \tag{27}$$
$$x_t = f(x_{t-1}) + g(x_{t-1})\epsilon_t \tag{28}$$

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - \Phi_1 B^m - \cdots - \Phi_P B^{Pm})(1 - B)^d (1 - B^m)^D N_{fts} = c + (1 + \theta_1 B + \cdots + \theta_q B^q)(1 + \Theta_1 B^m + \cdots + \Theta_Q B^{Qm})e_t \tag{29}$$

The environmental impact of current machines can be predicted as follows based on Eqs. (23)–(26):

$$I^{mfg} = \sum_r e_r^{raw} N_r + \sum_p e_p^{process} N_p + \sum_s e_s^{trans} N_s \tag{30}$$
$$\sum_{t=i}^{l} I^{usage} = \sum_{t=i}^{l}\sum_{s=1}^{z} e^{fuel} TS_{ts}^{N_f} + \sum_q \sum_{t=i}^{l}\sum_{s=1}^{z} e_q^{emission} ER_q TS_{ts}^{OH} \tag{31}$$
$$\sum_{t=i}^{l} I^{maint} = \sum_m \sum_{t=i}^{l}\sum_{s=1}^{z} e_m^{maint} N_m \left\lceil \frac{\max(TS_{ts}^{OH} - RC_m, 0)}{RC_m} \right\rceil \tag{32}$$
$$\sum_{t=i}^{l} I^{eol} = e_{used}^{eol} + \sum_m \sum_{t=i}^{l}\sum_{s=1}^{z} e_{replace}^{eol} N_m \left\lceil \frac{\max(TS_{ts}^{OH} - RC_m, 0)}{RC_m} \right\rceil \tag{33}$$

The second case, design for sustainability, is for the assessment of the new machines' sustainability when the target of environmental impact reduction should be applied to current machines due to new environmental regulations and enforcement. In this case, it is assumed that the new machines are upgraded versions of current machines. For example, new machines can improve the fuel efficiency with different materials or components. While these BOM (bill of materials) changes might increase the environmental impact of the manufacturing stage, the efficient fuel usage can reduce the environmental impact of the usage stage. As shown in Fig. 6, the main difference between the current machines and new machines is the availability of usage data (or usage model). The proposed method for the estimation of usage information is to use the improvement ratio which is defined as follows:

$$\delta_{N_f} = \frac{(N_f / W_{unit})_{\text{new machine}}}{(N_f / W_{unit})_{\text{current machine}}} \tag{34}$$
$$\delta_{OH} = \frac{(OH / W_{unit})_{\text{new machine}}}{(OH / W_{unit})_{\text{current machine}}} \tag{35}$$

where δ_{N_f} is the improvement ratio for the amount of fuel consumption, δ_{OH} is the improvement ratio for the operating hours, and W_{unit} is a unit of work. For example, if a new nutrient applicator can apply fertilizers with high precision and speed, these can be expressed as δ_{N_f} and δ_{OH} with the work unit of the square meter (m^2) from testing data. Then, the sensor data of current nutrient applicators can be used with the δ_{N_f} and δ_{OH} as follows for the environmental impact of the new machine:

$$I^{mfg} = \sum_r e_r^{raw} N_r + \sum_p e_p^{process} N_p + \sum_s e_s^{trans} N_s \tag{36}$$
$$\sum_{t=i}^{l} I^{usage} = \sum_{t=i}^{l}\sum_{s=1}^{z} e^{fuel} \delta_{N_f} TS_{ts}^{N_f} + \sum_q \sum_{t=i}^{l}\sum_{s=1}^{z} e_q^{emission} ER_q \delta_{OH} TS_{ts}^{OH} \tag{37}$$
$$\sum_{t=i}^{l} I^{maint} = \sum_m \sum_{t=i}^{l}\sum_{s=1}^{z} e_m^{maint} N_m \left\lceil \frac{\max(\delta_{OH} TS_{ts}^{OH} - RC_m, 0)}{RC_m} \right\rceil \tag{38}$$
$$\sum_{t=i}^{l} I^{eol} = e_{used}^{eol} + \sum_m \sum_{t=i}^{l}\sum_{s=1}^{z} e_{replace}^{eol} N_m \left\lceil \frac{\max\left(\delta_{OH} TS_{ts}^{OH} - RC_m, 0\right)}{RC_m} \right\rceil \tag{39}$$

The LCA result from Eqs. (36)–(39) estimates the environmental impact of the new machine. The result can also show whether the target of environmental impact reduction is satisfied. Otherwise, new design strategy should be explored. Note that the two design cases can be viewed as phases of a single design case, i.e., evaluation of current sustainability and redesign.

Numerical prediction tests for PUMLCA

In this section, a set of different data is tested to validate the prediction performance of PUMLCA. Due to the significance of environmental impact from the usage stage in LCA, the prediction accuracy of a time series usage model will play an important role for the estimation of environmental impact. The conventional method to model the usage stage is the constant rate method, which is the average of observations. The hypotheses are (1) the PUMLCA algorithm can provide a similar level of prediction accuracy to the constant rate method when data is constant with small random errors (i.e., steady-state processes), hereinafter *data 1*, (2) the PUMLCA can predict future values more accurately than the constant rate method when data has a trend, hereinafter *data 2*, (3) the automatic segmentation algorithm in PUMLCA can help to improve the predictive modeling when data has a trend and segments, hereinafter *data 3*, and (4) the PUMLCA algorithm can provide higher prediction accuracy than the constant rate method when prediction is required for specific periods within the whole prediction horizon.

Data sets (*data 1, 2, 3*) with monthly seasonal patterns were generated and the procedures are described in Section 'Data generation' for the hypotheses (1), (2) and (3). The three types of data sets were also used to test the hypothesis (4). In terms of the target of prediction, this study proposes to use not only the aggregated life cycle values (accuracy) but also the seasonal values of time series usage information (variance) because different time horizon scenarios can be tested. For example, monthly usage data is used to predict the next two-year values and the accumulated two-year values can be used to assess the environmental impact of life cycle as an accuracy measure. If the environmental impact of next quarter or specific periods within two years is required to be estimated, the accuracy of the predicted seasonal values (i.e., monthly values) will determine the quality of the analysis, which can be considered a variance measure. This is related to the fourth hypothesis. Therefore, the best model should provide good predictions of both values: high accuracy (aggregated life cycle values) and low variance (seasonal values).

As a prediction performance measure, mean absolute percentage error (MAPE) and mean absolute error (MAE) were used. Eqs. (40) and (41) show MAPE and MAE with the predicted values, $b_1, b_2, \ldots, b_m$ and the real values, $d_1, d_2, \ldots, d_m$. MAPE is scale-independent so that results from different data sets can be compared. However, by design, if the actual values are close to zero, MAPE cannot be defined. In this case, the scale-dependent measure, MAE, was used.

$$\text{Mean Absolute Percentage Error} = \frac{100\left(\left|\frac{b_1-d_1}{d_1}\right| + \cdots + \left|\frac{b_m-d_m}{d_m}\right|\right)}{m} \tag{40}$$

$$\text{Mean Absolute Error} = \frac{|b_1 - d_1| + \cdots + |b_m - d_m|}{m} \tag{41}$$

Note that based on MAPE and MAE, lower values of test results are preferable.

Throughout the numerical tests, only positive values were accepted as valid values. Negative values were set to zero. In order to handle non-negative data, one common method is the Box–Cox transformations (Hyndman and Athanasopoulos, 2013), which includes logarithms and power transformations. More theoretical discussions can be found in the literature (Hyndman et al., 2008).

Data generation

To test the first hypothesis, the following data generation procedure was applied: (1) a value from 100 to 1000 was randomly chosen using a random number generator for each month and (2) by adding a random error between −5 and 5 for each month, monthly data with seasonal patterns was generated for 16.5 years as shown in Table 8 in Appendix A. This is *data 1*, which does not contain a trend and segments.

For the second hypothesis, one more procedure was added from the procedure for *data 1*. After applying the first and second steps, 50 (i.e., a trend) was added for the next year values as shown in Table 9 in Appendix A (e.g., the column of January increases by 50). This is *data 2*, which contains a trend.

For the third hypothesis, after applying the first and second steps from the procedure for *data 1*, 100 (i.e., a trend) was added to the next year values. Then, eight consecutive monthly values starting from a random number were set to small numbers σ (i.e., segments) throughout the years as shown in Table 10 in Appendix A ($\sigma = 0$ in this test), which represents periods of no activity. This is *data 3*, which contains a trend and segments.

For each data, a total of 20 data sets were generated and tested. The first 7 years of data were used as training data and the remaining 9.5 years of data were used as test data as shown in Fig. 1.

Test results

The goal of this test is to construct a predictive model with the training data sets and predict future values (i.e., b_m in Eqs. (40) and (41)). The test data sets work as real values (i.e., d_m in Eqs. (40) and (41)). Table 1 shows the results of the data sets (*data 1, 2, 3*) in Section 'Data generation'.

First, for *data 1* (data without a trend and segments), since the data sets are designed to be constant with some mild randomness, the constant rate method showed good prediction performance for the accuracy measure. The PUMLCA algorithm

with both ETS (PUMLCA-ets) and ARIMA (PUMLCA-arima) also showed the similar level of accuracy and there is no significant difference between the constant rate method and PUMLCA (Mann–Whitney test, $\alpha = 0.05$, p-value = 0.95). For the variance measure, since the constant rate method took the average rate for each month, monthly predictions of the constant rate method showed much lower accuracy than those of PUMLCA (Mann–Whitney test, $\alpha = 0.05$, p-value = 0). This affects the prediction of the next quarter values (i.e., hypothesis 4) because lower monthly errors can give higher chances to predict specific periods with accuracy. For the next quarter values, the PUMLCA showed higher prediction accuracy (Mann–Whitney test, $\alpha = 0.05$, p-value = 0). Therefore, the PUMLCA algorithm can provide accurate prediction capabilities for aggregated life cycle values (accuracy), seasonal values (variance) and values for specific periods with *data 1* in comparison to the constant rate method.

Second, for *data 2* (data with a trend), the constant rate method showed poor prediction performance in terms of the accuracy measure. On the other hand, the PUMLCA algorithm with both ETS and ARIMA showed good prediction accuracy. There is no significant difference found between the real values and the results of PUMLCA-ets/arima (Mann–Whitney test, $\alpha = 0.05$, p-value = 0.29/0.78). For the variance measure, monthly predictions of the constant rate method showed much lower accuracy than those of PUMLCA (Mann–Whitney test, $\alpha = 0.05$, p-value = 0). This affects the prediction of the next quarter values. For the next quarter values, the PUMLCA showed higher prediction accuracy (Mann–Whitney test, $\alpha = 0.05$, p-value = 0). Therefore, the PUMLCA algorithm can provide accurate prediction capabilities for aggregated life cycle values (accuracy), seasonal values (variance) and values for specific periods with *data 2* in comparison to the constant rate method.

Third, for *data 3* (data with a trend and segments), the constant rate method and the ETS method without the automatic segmentation algorithm (ets-no seg) showed poor prediction performance in terms of the accuracy measure. On the other hand, the ARIMA method without the automatic segmentation algorithm (arimai-no seg) and PUMLCA-ets/arima showed strong prediction accuracy. However, Table 2 zooms in their prediction performances using MAE, and it can be seen that the errors from the ARIMA method without the automatic segmentation algorithm were much higher than the those from the PUMLCA method. Due to the importance of the usage stage, the errors from the ARIMA method without the automatic segmentation are not acceptable, and this shows that the automatic segmentation algorithm can enhance the prediction result. Out of 20 samples, the PUMLCA-ets/arima showed the best performance. For the next quarter values, the PUMLCA method with the automatic segmentation algorithm showed higher prediction accuracy. Therefore, the proposed segmentation algorithm can improve the predictive model of PUMLCA with *data 3*.

Overall, the PUMLCA method with the automatic segmentation algorithm provided better prediction performance than the constant rate method for various data sets which are simulated from the observation of real data. This prediction improvement of usage modeling will help to estimate the environmental impact of the product of interest more accurately. The example of the LCA with PUMLCA will be provided in the next section. The PUMLCA method could also provide prediction intervals while estimating a point forecast. For example, a point forecast of the next month is 1344 with the 80% prediction interval of [1330,1359]. The prediction interval can show the uncertainty of time series usage models.

Case study: agricultural machinery

Background

In this section, the proposed algorithm, predictive usage mining for life cycle assessment (PUMLCA), is demonstrated with a case study of agricultural machines: current and new machine. The machines have more than 15,000 parts and weigh more than 20,000 kg. The current machine was updated to have a 10% reduction of its environmental impact based on an improved fuel efficiency. This updated machine is called the new machine. The goal is to estimate the environmental impacts of the current and new machines in a real time horizon. Due to the data security issue, simulated data is used based on the

Table 1
Test results.

	Constant rate	ets-no seg	arima-no seg	PUMLCA-ets	PUMLCA-arima
Data 1, average MAPE					
Accuracy	0.75			0.08	0.14
Variance	65.58			0.76	0.79
Next quarter value	13.84			0.25	0.24
Data 2, average MAPE					
Accuracy	37.05			2.80	0.91
Variance	34.92			2.80	0.98
Next quarter value	22.06			0.74	0.29
Data 3, average MAE					
Accuracy	30,736	24,462	1612	166	154
Variance	636	313	225	2	2
Next quarter value	1979	1017	139	10	9

Table 2
MAEs over 20 data samples of *data 3*.

	1	2	3	4	5	6	7	8	9	10
arima-no seg	1870	3005	855	1478	2295	2382	592	2200	965	156
PUMLCA-ets	58	1061	64	48	311	292	17	237	101	34
PUMLCA-arima	145	1044	16	24	293	224	59	173	102	66
	11	12	13	14	15	16	17	18	19	20
arima-no seg	558	865	1870	1464	829	2829	1170	2826	1971	2060
PUMLCA-ets	96	70	540	9	102	122	66	3	48	48
PUMLCA-arima	57	0	322	147	119	80	35	64	47	66

observation of real data. Tables 3 and 4 show simulated seven-year monthly data for fuel consumption and operating hours after preprocessing the raw sensor data.

In this case study, time series usage models from the historical sensor data will be utilized to calculate the environmental impacts for up to 10–20 years. Since the first stage (i.e., data preprocessing in Section 'Data preprocessing') of PUMLCA is straightforward and simple, it was skipped in this section.

Seasonal period analysis

Instead of exploring all possible data representations (e.g., daily, weekly, quarterly, etc.), the focus was set on whether the simulated data showed a monthly seasonality. The periodogram was plotted using Eq. (3) with the condition of frequency greater than zero. The Periodogram shows that the maximum periodogram value can be achieved at the frequency of 0.0833 (i.e., period = 1/0.0833 = 12) for the fuel consumption data. Similarly, the operating hours data also indicate a period of 12.

Segmentation analysis

The automatic segmentation algorithm (Algorithm 1) was applied to the two data sets in Tables 3 and 4. As a penalty, the type I error of 0.05 was used for both data sets. First, for the fuel consumption data, a segment from February to August was identified as a shared segment since the same change points were detected (1, 8, 9, 10, 11, and 12 as seasonal time indexes) every year. Therefore, two segments were finally obtained, e.g., the shared segment (February–August) and the remaining segment (January, September–December). Second, for the operating hours data, the segment from January to August was

Table 3
Monthly representation of fuel consumption (ℓ) data.

Year	January	February	March	April	May	June	July	August	September	October	November	December
2007	9	0	0	0	0	0	0	2	600	3400	5000	250
2008	15	0	0	0	0	0	0	0	650	3410	5500	270
2009	17	0	0	0	0	0	0	0	660	3450	5550	280
2010	16	0	0	0	0	0	0	1	665	3370	5600	270
2011	14	0	0	0	0	0	0	1.5	660	3430	5650	275
2012	16	0	0	0	0	0	0	0	680	3500	5735	280
2013	17	0	0	0	0	0	0	2	700	3570	5800	285

Table 4
Monthly representation of operating hours (h) data.

Year	January	February	March	April	May	June	July	August	September	October	November	December
2007	1	0	0	0	0	0	0	0.2	35.2	100.6	152.3	15.1
2008	1.8	0	0	0	0	0	0	0	37.1	101.6	158.1	16.3
2009	2	0	0	0	0	0	0	0	38	105.3	159.3	17.8
2010	1.9	0	0	0	0	0	0	0.1	38.3	97.6	160.1	16.5
2011	1.7	0	0	0	0	0	0	0.2	38	103.5	162.2	17
2012	1.9	0	0	0	0	0	0	0	39	110.3	164.3	17.9
2013	2	0	0	0	0	0	0	0.22	41	115.2	165.2	18.2

identified as a shared segment. The same change points were detected (8, 9, 10, 11, and 12 as seasonal time indexes) every year. Therefore, two segments were finally obtained.

Time series analysis

The automatic forecasting algorithm in Section 'Automatic modeling of ETS and ARIMA' was applied to the original data sets (i.e., without segmentation) and the results of the automatic segmentation in Section 'Segmentation analysis'. Table 5 shows the results. For example, the original fuel consumption data is fitted as a seasonal AR model with a seasonal differencing and a drift using ARIMA. The first segment data (segment 1) shows a combination of seasonal AR and MA models without a drift. The second segment data (segment 2) shows only a seasonal differencing operation with a drift. The original fuel consumption data is also fitted as an additive error and seasonal component model using ETS. The first segment data shows an additive error and seasonal component model again. The second segment data shows an additive trend, multiplicative error and seasonal component model.

Predictive LCA

LCA for current machine

The PUMLCA-ets models (with two segments) of fuel consumption, N_{ft}, and operating hours, OH_t, in Table 5 were used as usage models of the agricultural machine. For predictive LCA, starting from January 2014, forecasts were built up to December 2024 (i.e., 10 years) and up to December 2034 (i.e., 20 years). For environmental impact calculation, Eco-Indicator 99 method (EI-99) (Goedkoop and Spriensma, 2001) was used, which is one of widely used methods in LCA and provides a single score (Point) from pre-defined damage categories such as human health, ecosystem quality, and resource.

In the manufacturing stage, the environmental impact was assumed as 12,000 Pt. In the usage stage, the density of diesel fuel was assumed as 0.85 kg/l and emission rates was given in Table 6. The idling and nonidling ratio (20%/80%) was calculated using averages of seven-year operating hours by work modes. In the maintenance stage, the assumptions on the replacement cycle of major parts and minor parts are as follows (Kwak and Kim, 2013): tires (3000 h), transmission (3000 h), hydraulic components (3000 h), engine (5000 h), axles (5000 h), and minor parts such as oils, greases, and filters (specified cycle). In the end-of-life stage, the following assumptions were made: steel (90% recycle and 10% landfill), iron (90% recycle and 10% landfill), and others (80% landfill and 20% incineration).

Based on Eqs. (30)–(33), a predictive LCA result of the current machine in the real time horizon (January 2014–December 2034) was estimated as shown in Fig. 7. The impact of the manufacturing stage was the same regardless of time horizons since it is a one-time event. On the other hand, the impacts of the usage, maintenance, and end-of-life stage were varied by time. Similar to previous LCA studies, the impact of the usage stage accounted for

Table 5
Results of time series analysis.

	ARIMA	ETS
Fuel consumption data		
Original	$(1-0.41B^{12})(1-B^{12})y_t = 1.53 + e_t$	$y_t = l_{t-1} + s_{t-12}$ $l_t = l_{t-1} + 0.06\epsilon_t$ $s_t = s_{t-12} + 10^{-4}\epsilon_t$
Segment 1 (February–August)	$(1+0.28B^{7})(1-B^{7})y_t = (1-0.28B^{4})e_t$	$y_t = l_{t-1} + s_{t-7}$ $l_t = l_{t-1} + 0.001\epsilon_t$ $s_t = s_{t-7} + 2\cdot 10^{-4}\epsilon_t$
Segment 2 (January, September–December)	$(1-B^{5})y_t = 7.42 + e_t$	$y_t = (l_{t-1} + b_{t-1})s_{t-5}$ $l_t = (l_{t-1} + b_{t-1})(1+0.395\epsilon_t)$ $b_t = b_{t-1} + 0.098(l_{t-1} + b_{t-1})\epsilon_t$ $s_t = s_{t-5}(1+10^{-4}\epsilon_t)$
Operating hours data		
Original	$(1-B^{12})y_t = (1+0.21B)e_t$	$y_t = l_{t-1} + s_{t-12}$ $l_t = l_{t-1} + 0.29\epsilon_t$ $s_t = s_{t-12} + 3\cdot 10^{-4}\epsilon_t$
Segment 1 (January–August)	$(1-B^{8})y_t = (1-0.67B)(1-0.64B^{8})e_t$	$y_t = l_{t-1} + s_{t-8}$ $l_t = l_{t-1} + 10^{-4}\epsilon_t$ $s_t = s_{t-8} + 0.03\epsilon_t$
Segment 2 (September–December)	$(1-B^{4})y_t = 0.38 + e_t$	$y_t = l_{t-1} + s_{t-4}$ $l_t = l_{t-1} + 0.12(l_{t-1} + s_{t-4})\epsilon_t$ $s_t = s_{t-4} + 0.88(l_{t-1} + s_{t-4})\epsilon_t$

Table 6
Assumptions on emission rates (g/h) (Kwak and Kim, 2013).

Type	Nonidling (80%)	Idling (20%)	Average
Nitrogen oxides (NOx)	372.73	143.16	326.82
Particulate matter (PM)	1.76	0.67	1.54
Carbon monoxide (CO)	23.84	9.16	20.9
Hydrocarbons (HC)	5.42	2.08	4.75
Sulfur dioxide (SO_2)	0.99	0.43	0.89
Carbon dioxide (CO_2)	150829.6	065427.83	133749.3

the majority of the environmental impact. The impact of the maintenance stage showed a big increase since major parts (engine and axles) were replaced after 10 years. It should be noted that the two usage models (PUMLCA and constant rate method) were used for the usage stage in order to show the impact of prediction accuracy in Section 'Numerical prediction tests for PUMLCA' (PUMLCA was also used for the maintenance and end-of-life stages). The data in this case study was similar to the third hypothesis in Section 'Numerical prediction tests for PUMLCA' (i.e., data with increasing trend and segments) so that it can be expected that the constant rate method would underestimate the impact (about 17,000 Pt over 20 years), which is greater than the impact of the manufacturing stage. If the data is quite constant, a similar result between PUMLCA and the constant rate method would be produced as seen in Section 'Numerical prediction tests for PUMLCA' (i.e., data without trend and segments). Furthermore, the top of Fig. 7 shows the 80% prediction intervals of the usage impact by PUMLCA. Unlike the constant rate method, PUMLCA can provide the uncertainty of its predictive model.

LCA for new machine

New machines were assumed to be designed based on the current machines with the target of 10% reduction of environmental impact over 20 years. It needs to utilize the usage data of the current machines with the improvement ratio, δ_{N_f} and δ_{OH} as shown in Fig. 6. Similar to the current machine, predictive LCA was conducted starting from January 2014 up to December 2024 (i.e., 10 years) and up to December 2034 (i.e., 20 years) with the EI-99 method.

In the manufacturing stage, the environmental impact was assumed to be increased to 14,500 Pt (20.8%) due to the additional power sources. The other assumptions of the usage, maintenance and end-of-life stage were similar to the current machine. The unit of work was the square meter (m^2) and the performance test was conducted to compare the new machine and the current machine. The improvement ratio for fuel consumption δ_{N_f} was 0.8 and the improvement ratio for operating hours δ_{OH} was 0.85.

Based on Eqs. (36)–(39), the predictive LCA result of the new machine in the real time horizon (January 2014–December 2034) was estimated as shown in Fig. 8.

Table 7 shows the comparison of the two LCA results of the current and new machine. Although the impact from the manufacturing stage was increased (20.8%) for the new machine, the total impact was reduced mainly from the usage stage. It should be noted that the result depends on the lifespan of machines. 8.4% of environmental impact reduction was expected

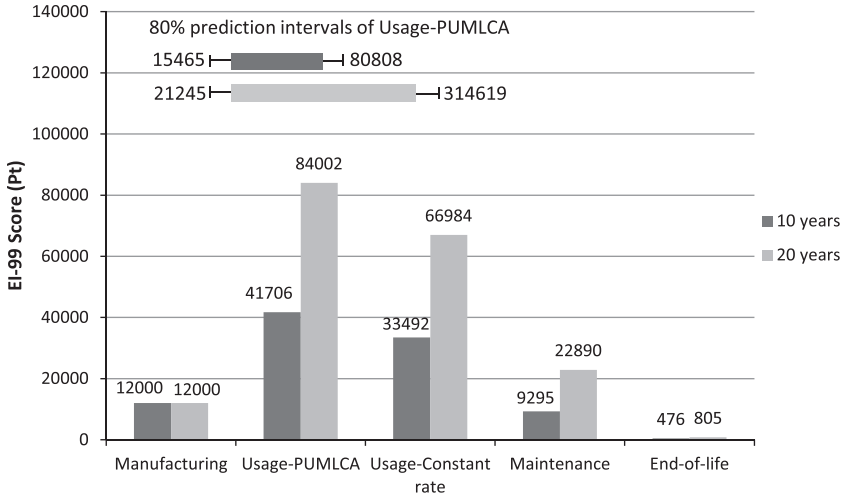

Fig. 7. Predictive LCA results for current machine.

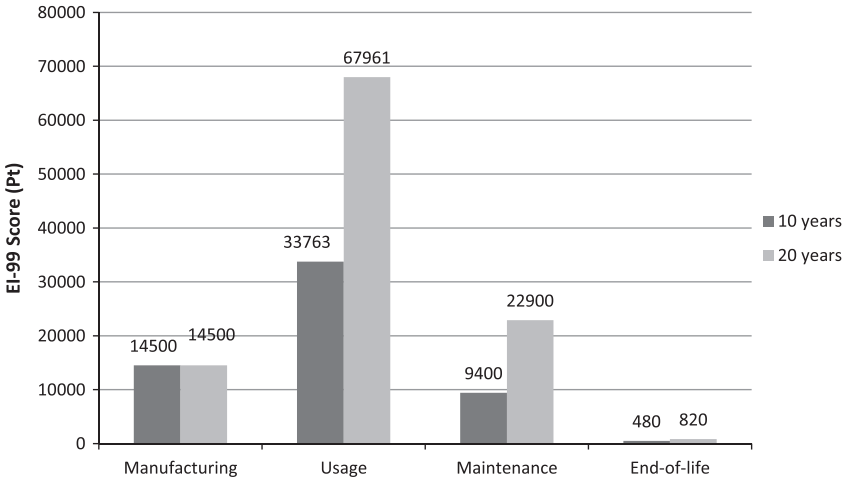

Fig. 8. Predictive LCA results for new machine.

Table 7
Comparison of current and new machines (EI-99, Pt).

	Manufacturing		Usage		Maintenance		End-of-life		Total	
	10 year	20 year	10 year	20 year	10 year	20 year	10 year	20 year	10 year	20 year
Current machine	12,000	12,000	41,706	84,002	9295	22,890	476	805	63,477	119,697
New machine	14,500	14,500	33,763	67,961	9400	22,900	480	820	58,143	106,181

for 10 years and 11.3% for 20 years, which satisfies the target of 10% reduction of environmental impact over 20 years. Sensitivity analysis can be applied to find the minimum values of the improvement ratio, δ_{N_f} and δ_{OH} to satisfy the target. In conclusion, the proposed algorithm, PUMLCA, captured usage patterns from large-scale sensor data with the automatic segmentation algorithm and time series analysis, and could assess environmental impact of a complex system in a real time horizon.

Closing remarks and future work

In this paper, the predictive usage mining for life cycle assessment (PUMLCA) algorithm is proposed to model the usage stage for the LCA of products. By defining usage patterns as trend, seasonality, and level from a time series of usage information, predictive LCA can be conducted in a real time horizon, which can provide more accurate results of LCA. Large-scale sensor data of product operation was analyzed to mine usage patterns and build a usage model for LCA. The PUMLCA algorithm includes handling missing and abnormal values, seasonal period analysis, segmentation analysis, time series analysis, and predictive LCA. In order to mine important usage patterns more effectively from a time series, the automatic segmentation algorithm is developed based on change point analysis.

The prediction performance test results with various data sets showed that the predictive model from the PUMLCA method can provide better prediction accuracy than the constant rate method. The automatic segmentation algorithm magnified important patterns and helped to predict future values more accurately.

Two different design problems were formulated to incorporate the usage model from the PUMLCA method in predictive LCA. The case study of agricultural machinery showed how to apply the PUMLCA method for the predictive LCA of complex systems. The environment impacts of both current machines and new machines could be estimated and compared.

In the future, various data sets from different products can be tested with the PUMLCA algorithm. The current model, which considers only a single type of machinery, can be extended to multiple types of machinery. In order to perform LCA with multiple types of machinery, hierarchical time series modeling and forecasting may be helpful (Hyndman et al., 2011).

Appendix A. Sample data sets in Section 'Numerical prediction tests for PUMLCA'

Tables 8–10 show the sample of *data 1*, *data 2*, and *data 3*.

Table 8
Sample of *data 1* for hypotheses 1 and 4.

Year	January	February	March	April	May	June	July	August	September	October	November	December
1	470	538	544	669	232	911	747	353	909	980	133	213
2	475	540	545	672	231	913	742	354	909	982	130	218
3	475	542	544	670	234	908	747	354	914	985	129	215
4	466	539	547	671	229	919	745	350	906	975	135	216
5	473	534	548	674	232	913	748	358	913	984	135	214
6	474	539	539	668	232	911	747	349	908	983	132	208
7	471	541	548	667	232	913	748	353	912	982	137	214
8	473	543	545	666	229	907	748	354	911	980	136	217
9	467	536	542	670	229	911	745	355	907	975	138	211
10	466	537	544	674	235	914	743	355	910	979	136	217
11	468	536	543	673	230	909	749	349	909	982	129	215
12	472	542	542	665	222	908	750	351	908	976	132	208
13	466	541	545	664	229	916	746	351	905	977	132	218
14	473	542	539	667	229	912	742	354	908	977	133	217
15	474	538	541	664	228	914	748	349	905	984	133	209
16	473	533	549	674	232	911	751	356	909	979	135	212
17	467	534	539	672	234	915						

Table 9
Sample of *data 2* for hypotheses 2 and 4.

Year	January	February	March	April	May	June	July	August	September	October	November	December
1	975	872	965	976	799	449	681	169	399	728	614	725
2	1024	921	1010	1029	845	500	733	219	455	779	669	772
3	1077	973	1061	1070	893	549	786	271	502	828	713	823
4	1123	1022	1119	1129	940	605	832	312	549	872	765	871
5	1174	1077	1160	1179	991	659	885	365	600	928	813	923
6	1224	1117	1210	1224	1040	701	930	421	658	974	870	975
7	1273	1176	1268	1275	1095	751	978	462	698	1030	913	1025
8	1326	1220	1309	1325	1139	808	1029	522	751	1073	963	1079
9	1381	1271	1359	1379	1197	854	1078	567	805	1130	1011	1131
10	1427	1321	1419	1421	1248	899	1128	616	857	1179	1063	1181
11	1481	1367	1468	1472	1299	950	1180	671	900	1230	1117	1225
12	1526	1419	1515	1526	1340	1006	1229	712	953	1278	1162	1278
13	1575	1469	1561	1569	1393	1058	1278	769	1005	1328	1215	1328
14	1629	1527	1618	1625	1447	1099	1337	821	1056	1371	1268	1380
15	1677	1575	1667	1670	1495	1155	1378	872	1102	1420	1320	1423
16	1723	1623	1714	1722	1540	1209	1429	933	1153	1469	1372	1471
17	1770	1671	1760	1776	1592	1258						

Table 10
Sample of *data 3* for hypotheses 3 and 4.

Year	January	February	March	April	May	June	July	August	September	October	November	December
1	σ	σ	σ	σ	σ	σ	155	129	643	313	σ	σ
2	σ	σ	σ	σ	σ	σ	257	233	746	409	σ	σ
3	σ	σ	σ	σ	σ	σ	355	333	848	518	σ	σ
4	σ	σ	σ	σ	σ	σ	452	429	944	610	σ	σ
5	σ	σ	σ	σ	σ	σ	558	525	1038	710	σ	σ
6	σ	σ	σ	σ	σ	σ	654	632	1141	813	σ	σ
7	σ	σ	σ	σ	σ	σ	752	734	1242	909	σ	σ
8	σ	σ	σ	σ	σ	σ	855	827	1344	1012	σ	σ
9	σ	σ	σ	σ	σ	σ	958	928	1445	1117	σ	σ
10	σ	σ	σ	σ	σ	σ	1053	1025	1542	1214	σ	σ
11	σ	σ	σ	σ	σ	σ	1160	1124	1643	1317	σ	σ
12	σ	σ	σ	σ	σ	σ	1253	1231	1743	1410	σ	σ
13	σ	σ	σ	σ	σ	σ	1354	1328	1839	1510	σ	σ
14	σ	σ	σ	σ	σ	σ	1450	1425	1943	1616	σ	σ
15	σ	σ	σ	σ	σ	σ	1553	1534	2044	1711	σ	σ
16	σ	σ	σ	σ	σ	σ	1656	1629	2143	1808	σ	σ
17	σ	σ	σ	σ	σ	σ						

References

Choi, B.C., Shin, H.S., Lee, S.Y., Hur, T., 2006. Life cycle assessment of a personal computer and its effective recycling rate (7 pp). Int. J. Life Cycle Assess. 11, 122–128.

Collet, P., Lardon, L., Steyer, J.P., Hélias, A., 2014. How to take time into account in the inventory step: a selective introduction based on sensitivity analysis. Int. J. Life Cycle Assess. 19, 320–330.

Finnveden, G., Hauschild, M.Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., Suh, S., 2009. Recent developments in life cycle assessment. J. Environ. Manage. 91, 1–21.

Goedkoop, M., Spriensma, S., 2001. The Eco-Indicator 99: A Damage Oriented Method for Life Cycle Impact Assessment. Annex Report. Pre Consultant, B.V. Amersfoort, The Netherlands. <Http://www.pre-sustainability.com>.

Guinée, J., 2002. Handbook on Life Cycle Assessment: Operational Guide to the ISO Standards. Eco-Efficiency in Industry and Science. Springer.

Hyndman, R., Athanasopoulos, G., 2013. Forecasting: Principles and Practice. <Http://otexts.org/fpp/> (accessed January 2014).

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. J. Stat. Softw. 27, 1–22.

Hyndman, R., Koehler, A., Ord, J.K., Snyder, R., 2008. Forecasting with Exponential Smoothing: The State Space Approach. Springer-Verlag, Berlin, Heidelberg.

Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L., 2011. Optimal combination forecasts for hierarchical time series. Comput. Stat. Data Anal. 55, 2579–2589.

Jackson, T., 2010. Analyzing seasonal time series with periodic low volumes. In: Proceedings of International Symposium on Forecasting, San Diego, USA.

Keogh, E., Chu, S., Hart, D., Pazzani, M., 2004. Segmenting time series: a survey and novel approach. Data Min. Time Ser. Databases 57, 1–21.

Killick, R., Eckley, I.A., 2011. Changepoint: An R Package for Changepoint Analysis. R Package Version 0.5.

Killick, R., Fearnhead, P., Eckley, I.A., 2012. Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. 107, 1590–1598.

Kwak, M., 2012. Green Profit Design for Lifecycle. Ph.D. Thesis, University of Illinois at Urbana-Champaign.

Kwak, M., Kim, H., 2013. Economic and environmental impacts of product service lifetime: a life-cycle perspective. In: Meier, H. (Ed.), Product-Service Integration for Sustainable Solutions, Lecture Notes in Production Engineering. Springer, Berlin, Heidelberg, pp. 177–189.

Kwak, M., Kim, L., Sarvana, O., Kim, H.M., Finamore, P., Hazewinkel, H., 2012. Life cycle assessment of complex heavy duty equipment. In: ASME International Symposium on Flexible Automation (ISFA2012), St. Louis, USA.

Lee, J., Cho, H.J., Choi, B., Sung, J., Lee, S., Shin, M., 2000. Life cycle assessment of tractors. Int. J. Life Cycle Assess. 5, 205–208.

Levasseur, A., Lesage, P., Margni, M., Deschênes, L., Samson, R., 2010. Considering time in LCA: dynamic LCA and its application to global warming impact assessments. Environ. Sci. Technol. 44, 3169–3174.

Li, T., Liu, Z.C., Zhang, H.C., Jiang, Q.H., 2013. Environmental emissions and energy consumptions assessment of a diesel engine from the life cycle perspective. J. Cleaner Prod. 53, 7–12.

Ma, J., Kim, H.M., 2014. Continuous preference trend mining for optimal product design with multiple profit cycles. J. Mech. Des. 136, 061002.

Ma, J., Kwak, M., Kim, H.M., 2014. Demand trend mining for predictive life cycle design. J. Cleaner Prod. 68, 189–199.

Memary, R., Giurco, D., Mudd, G., Mason, L., 2012. Life cycle assessment: a time-series analysis of copper. J. Cleaner Prod. 33, 97–108.

Reap, J., Roman, F., Duncan, S., Bras, B., 2008a. A survey of unresolved problems in life cycle assessment. Part 1: Goal and scope and inventory analysis. Int. J. Life Cycle Assess. 13, 290–300.

Reap, J., Roman, F., Duncan, S., Bras, B., 2008b. A survey of unresolved problems in life cycle assessment. Part 2: Impact assessment and interpretation. Int. J. Life Cycle Assess. 13, 374–388.

Rebitzer, G., Ekvall, T., Frischknecht, R., Hunkeler, D., Norris, G., Rydberg, T., Schmidt, W.P., Suh, S., Weidema, B., Pennington, D., 2004. Life cycle assessment: Part 1: framework, goal and scope definition, inventory analysis, and applications. Environ. Int. 30, 701–720.

Shumway, R., Stoffer, D., 2011. Time Series Analysis and Its Applications: With R Examples. Springer Texts in Statistics. Springer.

Sullivan, J.L., Cobas-Flores, E., 2001. Full vehicle LCAs: a review. In: Proceedings of the 2001 Environmental Sustainability Conference and Exhibition, Graz, Austria, pp. 99–114.

Telenko, C., Seepersad, C.C., 2014. Probabilistic graphical modeling of use stage energy consumption: a lightweight vehicle example. J. Mech. Des. 136, 101403.