# *Chapter 8*

# Autoregressive Bayesian Networks for Information Validation and Amendment in Military Applications

Pablo Ibargüengoytia, Javier Herrera-Vega, Uriel A. García, L. Enrique Sucar, and Eduardo F. Morales

## Contents

## Introduction

Data integration and data cleansing are particularly relevant for military applications where trustable data can make a difference in life-threatening conditions. One of the aims in military data management is to have information superiority, which largely depends on the ability to have the right data at the right time. The right data not only means relevant information but also trustable information content that is used when the decision-making processes are made. There is a large number of military applications that depends on data obtained from different sources that need integration and cleansing, such as military surveillance and security domains.

An important and current application of data management is in the fight against terrorism. The key idea is to investigate and understand criminal behavior based on historical data extracted from pasts events. Large databases are being constructed in order to discover behavioral patterns that permit predicting future attacks.

In February 2017, the BBC published an article that relates the use of Facebook to detect terrorist activities using machine-learning algorithms. When using Twitter, a database can be formed with different attributes like text content, user ID, other tagged individuals, timestamps, the language device used, and the user's location.

Caruso (2016) describes how Facebook and Twitter have special teams that detect terrorist activity on their social media and remove individuals or groups associated with terrorist activities. In 2016, Twitter suspended 125,000 accounts with links to ISIS.

Tutun et al. (2017) utilize historical data of patterns followed by 150,000 terrorist attacks from 1970 to 2015. The authors comment that terrorists have learned that using social media risks detection and hence use alternatives. The analysis of databases includes 140,000 incidents considering approximately 75 features or attributes that characterize the criminal behavior.

This chapter proposes a novel and robust mechanism for information validation and amendment in databases where certainty in information is critical.

In data-validation problems, missing data requires estimation and, if inaccurate, requires rectification. In both cases, anomaly detection and information reconstruction are necessary. Additionally, some contextual information may be needed, whether available from the same or a complementary data source. Given the critical importance that access to trustable data has to many industries, including the military, it is no surprise that data validation methods have been

thoroughly researched across many of the more common issues: outliers (Abraham and Box, 1979; Balke, 1993; Hoo et al., 2002; Muirhead, 1986; Peng et al., 2012; Tsay, 1988; Walczak, 1995), sudden changes also referred to as innovation outliers (Abraham and Box, 1979; Balke, 1993; Marr and Hildreth, 1980; Muirhead, 1986; Tsay, 1988; Sato et al., 2006), rogue values (Herrera-Vega et al., 2012; Ibargüengoytia et al., 2006), and missing data (Dempster et al., 1977; Lamrini et al., 2011; Vagin and Fomina, 2011). Any of these deviations from regular data behavior may actually be due to exceptional circumstances and do not necessarily represent inaccurate information. In those cases, amendment is not necessary and the unusual data leads to actions such as raising an alert. But following the confirmation of inaccurate information, whatever the cause, the failing data has to be reconstructed, and for such purpose, it can be treated as missing.

Missing data refers to the problem where a gap in the information exists. Etiology is varied; it may be a missing sample, a necessity of out-of-boundaries inference, or the demand for a resampling, among others. In situations of missing information, interpolation/extrapolation techniques (Lancaster and Salkauskas, 1986) have dominated the scene. But interpolation is not the only option. When the available information comes from a single source with a certain temporal structure, then classical time-series modelling (Chatfield, 2004) such as Autoregressive Moving Average (ARMA) or Integrated Autoregressive Moving Average (ARIMA) has also been employed. Both approaches are appropriate for isolated data series and capitalize on within-variable information. When richer contextual information from a number of additional variables is available, a range of alternative techniques should be considered to exploit the complementary knowledge. These multivariate techniques may afford a reconstruction of the missing datum in terms of the nearest neighbor (Vagin and Fomina, 2011), self-organizing maps (Lamrini et al., 2011), or probabilistic graphic networks (Ibargüengoytia et al., 2013a), among others. These multivariate approaches have in common the exploitation of adjacent variables, often at the cost of ignoring any signal own information. In contrast with the rich AQ 2 literature available on different validation methods, the decision of when to choose one particular reconstruction strategy over another has been scarcely investigated. Little is known about when the data-set characteristics will favor the application of one technique over another. In such uncertain scenarios, a method that utilizes both sources of information, the signal-internal information and the related information present in the repository, may represent a compromise of the advantages of different approaches while alleviating the process of picking the best-suited data estimation approach.

Ideally, both the signal-internal information and the related information present in the repository should be taken into account for estimating the missing information, but this is an oversimplification. In every case, the weight given to the in-variable information, and the information from other variables should be reconsidered. In our research work preceding this chapter (Ibargüengoytia et al., 2013b), we showed how the performance of different data estimation approaches vary as the

scenario of available information exhibits different properties, and more precisely, how this is dictated by the variable autoregressive order and its dependency on the additional known variables. In that previous work, a concomitant contribution was the proposal of a new model, the Autoregressive Bayesian Network (AR-BN), that balanced its output, aiming to perform robustly across a wide range of scenarios.[1]

# Methods

In this section, a description of some available methods to complete databases is included. Later, a more complete description of the proposal of this chapter, namely the Autoregressive Bayesian Network, is included. A mathematical formulation of the problem of incomplete data can be found in Dempster et al. (1977).

## *Interpolation*

Interpolation is a large family of models in which the value of a variable at some sampling location is estimated from neighbor observations. In general, the semantics of the sampling location is irrelevant for the model itself other than setting, which are the neighbor samples and how far they may be from the questioned sampled location. When the sampling location is within other observed locations, then these models are referred to as interpolation, and when beyond, then they are referred to as extrapolation. Traditionally, interpolation has been an easier guess than extrapolation.

### *Linear Interpolation*

Perhaps the simplest interpolation approach is linear interpolation, by which, assuming that the function is locally linear, that is, the approximation using only the Jacobian system is considered sufficient, the value of the variable at the new location is given by the line crossing its two nearest-known neighbor observations. Let $s_t$[2] be the targeted new sampling location and discretely let $s_{t-1}$ and $s_{t+1}$ be the immediate previous and next neighbor locations on which observations $X_{t-1}$ and $X_{t+1}$ have already been made. The estimated value $X_t$ is given by Equation 8.1:

$$X_t = X_{t-1} + (s_t - s_{t-1}) \frac{X_{t+1} - X_{t-1}}{s_{t+1} - s_{t-1}} \tag{8.1}$$

[1] This chapter is an extension of the work presented by this group at The Eighth International Conference on Systems (ICONS) 2013 (Ibargüengoytia et al., 2013b), with emphasis on the description of the AR-BN model.

[2] Without any temporal semantics associated.

### *Spline Interpolation*

Spline interpolation is a more sophisticated approach in which a differentiable curve is built in a piecewise manner between an arbitrary number of neighbor observations supporting the curve definition. The curve is expected to be differentiable at all supporting points and thus derivatives at those points also ought to be known. In the simplest case, when only two supporting points $s_{t-1}$ and $s_{t+1}$ and associated observations $X_{t-1}$ and $X_{t+1}$ are taken (the first derivatives at those points $X'_{t-1}$ and $X'_{t+1}$ can be estimated from using further subsequent neighbors), a third-order polynomial can be written as:

$$X_t = (1-h)X_{t-1} + hX_{t+1} + h(1-h)(a(1-h)+bh) \tag{8.2}$$

where:

$$h = \frac{s_t - s_{t-1}}{s_{t+1} - s_{t-1}}$$
$$a = X'_{t-1}(s_{t+1} - s_{t-1}) - (X_{t+1} - X_{t-1})$$
$$b = -X'_{t+1}(s_{t+1} - s_{t-1}) + (X_{t+1} - X_{t-1})$$

As the earlier system is underdetermined, second derivatives are commonly required to match the sampling locations to complete the system. Other polynomials can be constructed (de Boor, 1978), but they are beyond the scope of this chapter.

## *Autoregressive Models*

Time-series analysis has traditionally focused on estimating future values of a variable that has (or is assumed to have) a certain dynamic, for example, general trend, seasonalities, and so on. Autoregressive models are perhaps the simplest of time-series models in which the next observation is derived from a linear combination of preceding observations. The general autoregressive model of order $n$—denoted AR($n$)—is defined in Equation 8.3:

$$X_t = c + \sum_{i=1}^{n} \alpha_i X_{t-i} + \epsilon_t \tag{8.3}$$

where $\alpha_{i|i=1\ldots n}$ are the model parameters, $c$ is a constant, and $\epsilon_t$ is noise. Note the assumed temporal semantics in contrast to interpolation, but beware that from an abstract point of view, it remains a discrete relation of ordering among the sampling locations. Hence, in an offline repository, "future" data may also be available and can be easily incorporated as in Equation 8.4:

$$X_t = c + \sum_{i=1}^{n} \alpha_i X_{t-i} + \sum_{j=1}^{m} \beta_j X_{t+j} + \epsilon_t \tag{8.4}$$

where $\alpha_{i|i=1\ldots n}$ and $\beta_{j|j=1\ldots m}$ are the AR($n$, $m$) model parameters.

## Probabilistic Modeling

Probabilistic models exploit the laws of probability to estimate the most likely outcome of some event, that is, location, defined over the probability space (Sucar, 2015). Like interpolation and autoregressive models, probabilistic models are also a large family that traditionally have been well-suited under uncertainty. Among them, a Bayesian network is a directed acyclic graph (DAG) representing the joint probability distribution of all variables in a domain (Pearl, 1988). Bayesian networks use the Bayes theorem that relates hypotheses and evidence and makes relations among variables graphically explicit, as can be seen in Figure 8.1. The graphical companion is not superfluous. The topology of the network conveys direct information about the dependency between the variables. The structure of the graph represents which variables are conditionally independent given another variable. The variable at the end of an arc end (variable $E$) is probabilistically dependent on the variable at the origin of the arc (variable $H$).

Thus, obtaining values of the evidence $E$, Bayesian networks calculate the probability of hypothesis $H$ given the evidence. This corresponds to Bayes' theorem where the computation of $P(H \mid E)$ is calculated using $P(E \mid H)$ and $P(H)$ as per Equation 8.5.

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \tag{8.5}$$

The knowledge in a process using Bayesian networks can be represented with two elements: (1) the structure of the network, and (2) the parameters $P(E \mid H)$, that is, the conditional probability tables, and $P(H)$. These parameters are learned from observations when available. In the application of completing databases, the parameter $P(\text{missingvalues} \mid \text{relatedvalues})$ can be calculated using $P(\text{relatedvalues} \mid \text{missingvalues})$ if knowledge is available in these databases, that is, with complete historical data-set of the process.

AQ 3

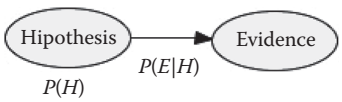



**Figure 8.1 Elemental Bayesian Network: Structure and parameters.**

Knowledge about a system represented as a Bayesian network can be used to reason about the consequences of specific input data by what is called probabilistic reasoning. This consists of assigning a value to the input variables and propagating their effect through the network to update the probability of the hypothesis variables. The updating of the certainty measures is consistent with probability theory based on the application of Bayesian calculus and the dependencies represented in the network. Several algorithms have been proposed for this probability propagation (Pearl, 1988).

Bayesian networks can use historical data to acquire knowledge but may additionally assimilate domain experts' input. One of the advantages of using Bayesian networks is the three forms to acquire the required knowledge. First, the participation of human experts in the domain is quite effective; they can explain the dependencies and independencies between the variables and also may calculate the conditional probabilities. Second, there is a great variety of automatic-learning algorithms that utilize historical data to provide the structure and the conditional probabilities corresponding to the process where data was obtained. A combination of the previous two is the third approach, that is, using an automatic-learning algorithm that allows for the participation of human experts in the definition of the structure.

## *Dynamic Bayesian Networks*

Plain Bayesian Networks (BN) consider only static situations of a domain. Time is not considered, and the calculation made on the hypothesis nodes consider only current values of the evidence nodes. The databases relevant in data validation in this chapter are usually time series where values are obtained in discrete intervals of time. Dynamic Bayesian Networks (DBN) are an attempt to add the temporal dimension into the BN model (Dean and Kanazawa, 1989; Mihajlovic and Petkovic, 2001). Often a DBN incorporates two models: an initial net $B_0$, learned using information at time 0, and the transition net $B_{\rightarrow}$, learned with the rest of the data as illustrated in Figure 8.2. Together, $B_0$ and $B_{\rightarrow}$ constitute the DBN (Koller and Friedman, 2009). An important assumption is made for DBNs: The process is assumed to be Markovian, that is, the future is conditionally independent of the past given the present. This assumption allows the DBN to use only the previous time-stage information in order to obtain the next stage.

A DBN can be unfolded over as many stages as necessary, and the horizontal structure can change from stage to stage. The resulting network is highly expressive but often unnecessarily complicated. Alternatives have been proposed to reduce this complexity like the Temporal Nodes Bayesian Networks (Herrera-Vega et al., 2012). In data-sets arising from physical processes, statistical dependencies among variables can be expected to be stable across time. That is, if two variables, $X$ and $Y$, are statistically dependent at time $t_i$, they will likely also be statistically dependent at time $t_{i+j}$ for any arbitrary samples $i$ and $j$, and similar reasoning can be made for independencies. This implies that the process is time-invariant, which can be

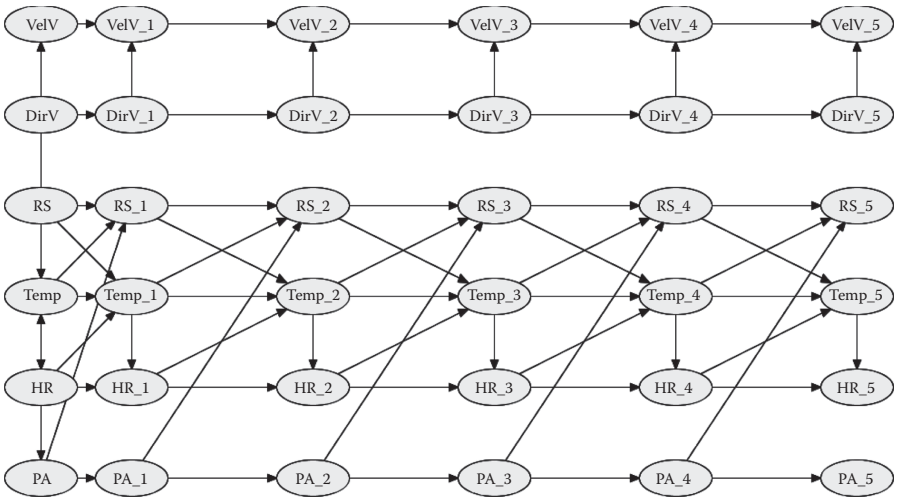



**Figure 8.2 Dynamic Bayesian Network.**

exploited to simplify the model representation, as only the initial and transition networks are required.

## *Autoregressive Bayesian Networks (AR-BN)*

Autoregressive Bayesian Networks are a simplified variant of DBNs. They incorporate the temporal dimension by observing time-shifted versions of the variables, whether past or future. Conceptually, they can be regarded as bringing an autoregressive model AR($n$, $m$) to the BN domain.

Suppose a data-set with some dynamics of interest. Figure 8.3 illustrates the proposed probabilistic model. Variable $X$ represents the variable to be estimated,

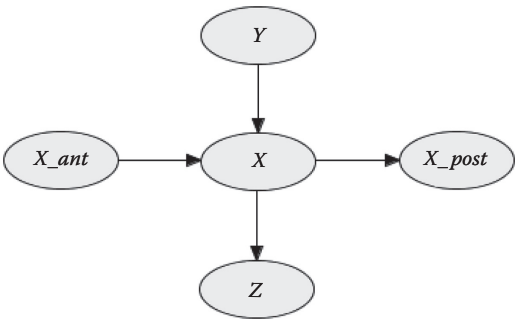



**Figure 8.3 Dynamic probabilistic model proposed for data estimation. The structure can be enriched with other time-shifted versions of $X$, $Y$, and $Z$ as appropriate.**

variables $Y$ and $Z$ represent pieces of Bayesian network corresponding to all the related variables to $X$. *X_post* represents the value of variable $X$ at the time $t + k$, and *X_ant* represents the value of variable $X$ at the time $t - k$, although for simplicity in this chapter we will take $k = 1$.

This proposed model represents a dynamic model that provides accurate information for estimating the variable in two senses: first, by using related information identified by automatic-learning algorithms or experts in the domain, or both; and second, by using information of the previous and incoming values. This information includes the change rate of the variable according to the history of the signal.

In this approach, the horizontal (inter-stage) topology of the network is fixed. The persistency arcs among a variable and its shifted versions are enforced, whereas those between different variables at different stages are forbidden.

## *Estimating Missing Data from Incomplete Databases Using AR-BN*

The proposed procedure for estimating missing data from incomplete databases is in Algorithm 8.1. The first 4 steps build the model, and the last 3 propagate knowledge to estimate data holes.

Suppose a time-series of three variables. Following Algorithm 8.1, a structure of the static version is obtained in step 3, as shown in Figure 8.4. In step 4, the network is extended and completes an AR-BN as shown in Figure 8.5.

**Algorithm 8.1 Estimation of Missing Data**

| |
|---|
| 1 : Obtain a complete data-set that includes information from the widest operational conditions of the target process. |
| 2 : Clean the outliers and discretize the data set. |
| 3 : Utilize a learning algorithm that produces the static Bayesian Network relating all the variables in the process. During the learning process, a complete train set with data from all variables is needed, as indicated in step 1. |
| 4 : Modify the static model to include previous and posterior values of every variable. |
| 5 : For all registers in an incomplete database, if one value is missing, instantiate the rest of the nodes in the model. |
| 6 : Propagate to obtain a posterior probability distribution of the missing value given the available evidence. |
| 7 : Return the estimated value with the value of the highest probability interval, or calculate the expected value of the probability distribution. |

AQ 4

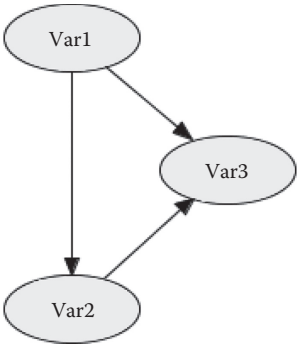



**Figure 8.4 Static Bayesian Network version of data-set 2.**

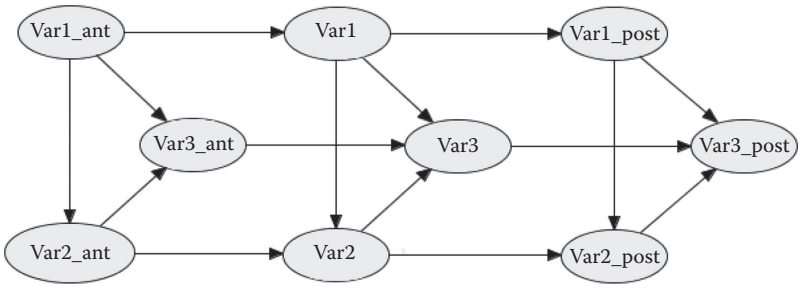



**Figure 8.5 Autoregressive Bayesian Network proposed for data estimation for data-set 2.**

Notice that the creation of the structure, as in Figure 8.5, requires the calculation of parameters. These parameters are calculated, as mentioned in step 1 of Algorithm 8.1, using a complete data-set of the operational conditions of the target process.

## *Error Metrics for the Estimation of Missing Values*

In order to establish the accuracy of the estimation of the missing values, the following error metrics were computed (Osman et al., 2001). Let $E_i$ be the relative deviation of an estimated value $x_i^{est}$ from an experimental value, $x_i^{obs}$:

$$E_i = \left[ \frac{x_i^{obs} - x_i^{est}}{x_i^{obs}} \right] \times 100 i = 1,2,\ldots,n \tag{8.6}$$

with $n$ being the number of missing data.

- **Root Mean Square Error:**

$$E_{rms} = \left[ \frac{1}{n} \sum_{i=1}^{n} E_i^2 \right]^{1/2} \tag{8.7}$$

- **Average Percent Relative Error:**

$$E_r = \frac{1}{n} \sum_{i=1}^{n} E_i \tag{8.8}$$

- **Average Absolute Percent Relative Error:**

$$E_a = \frac{1}{n} \sum_{i=1}^{n} | E_i | \tag{8.9}$$

- **Minimum and Maximum Absolute Percent Relative Error:**

$$E_{min} = min_{i=1}^{n} | E_i | \tag{8.10}$$

$$E_{max} = max_{i=1}^{n} | E_i | \tag{8.11}$$

These metrics will be used in the experiments conducted and discussed in section "Experiments and Results".

## *Data Characterization*

By analyzing the data, we can gain insight into the difficulty of validating and amending the data-sets, as well as which method to complete the missing data could be more appropriate. To extract descriptive parameters that help us to know more about the behavior of the data-sets, we have characterized them according to the methods described as follows:

- **Principal Component Analysis (PCA):** This technique, developed by Hotelling (1933), analyzes a data-set composed by intercorrelated variables and extracts the relevant information in the data. Then, this is represented as a set of orthogonal variables called principal components that correspond with the maximum variance. Data dimensionality is reduced by removing the components with less variance. The resulting variables represent the intrinsic dimensionality of the original data-set.
- **Intrinsic dimensionality:** The algorithm of Fukunaga and Olsen (1971) aims to look at local characteristics of the data distribution, establishing small subregions around each variable and, by the Karhunen-Love expansions for these subregions, determine the intrinsic dimensionality of the data.

- **Pearson correlation:** This is a well-known method to measure the linear dependence between two variables. Values +1 and −1 represent a linear dependence, and 0 indicates a nonlinear relation. The Pearson correlation coefficient is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{8.12}$$

- **Akaike Information Criterion (AIC):** The AIC is a model selection method (Akaike, 1969) defined as:

$$AIC = 2 * N \log L + 2 * m \tag{8.13}$$

where $m$ is the number of estimated parameters, and $N\log L$ is the log-likelihood. This method selects a model that minimizes the distance between the model and the truth. In autoregressive models, with Akaike's method, we select the order for which Equation 8.13 attains its minimum as a function of $m$ (Shibata, 1976).

- **Kwiatkowski–Phillips–Schmidt–Shin (KPSS):** This method tests the null hypothesis that a time-series is trend-stationary against the alternative hypothesis that it is a nonstationary process (Kwiatkowski et al., 1992). Briefly, the KPSS breaks the time-series in three parts to construct a model (Equation 8.14) consisting of: a deterministic trend ($\beta t$), a random walk ($r_t$), and a stationary error ($\varepsilon_t$):

$$x_t = r_t + \beta t + \varepsilon \tag{8.14}$$

A least-squares regression is performed to fit the original data and the model. Finally, the data is considered stationary if the term ($r_t$) is constant.

## Experiments and Results

This section describes the set of experiments conducted for the comparison of performance between different methods on different data-sets.

### *Characterization of the Datasets for the Experiments*

Simulations were carried out to reconstruct missing data from 2 different industrial data-sets of different natures (variables have been enumerated for confidentiality).

The first data-set comprises 10 variables. It corresponds to a manufacturing process. The second data-set comprises 3 variables. It corresponds to an energy domain. Intrinsic dimensionality of the data-sets as found by Principal Component Analysis (PCA) is 7 and 1 respectively (99% of variance included). For the data-set 2, the scale of one of the variables is 5 orders of magnitude larger than the remaining 2 variables. Hence, the global intrinsic dimensionality is perceived to be 1 by PCA, but local dimensionality of the dataset remains 3, which can be determined by Fukunaga and Olsen's algorithm (Chatfield, 2004). The pairwise Pearson correlations among variables for the data-sets in Figure 8.5 hint about the dependencies among variables. The variables autoregressive order n was estimated using the Akaike Information Criterion (Kwiatkowski et al., 1992), providing an indication of the signal-own predictability. The autoregressive orders found with this criterion are summarized in Table 8.1. Stationarity of the time-series was estimated using the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test for stationarity and is summarized in Table 8.2 (Figure 8.6).

**Table 8.1 Autoregressive Orders as Calculated with the Akaike Information Criterion**

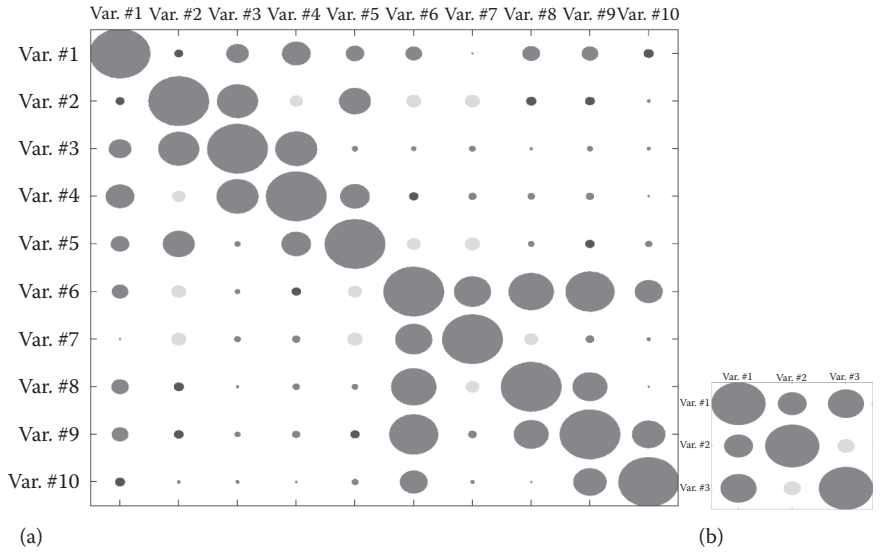
| *Var.#* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset 1 | 2 | 2 | 1 | 2 | 2 | 9 | 7 | 9 | 9 | 9 |
| Dataset 2 | 25 | 1 | 25 | | | | | | | |

**Table 8.2 Kwiatkowski–Phillips–Schmidt–Shin (KPSS) Stationarity Tests**

| *Var.#* | *Dataset 1* | *Dataset 2* |
|---|---|---|
| 1 | $p < 0.01$[a] | $p = 0.01$[b] |
| 2 | $p < 0.01$[a] | $p = 0.014$[b] |
| 3 | $p < 0.01$[a] | $p = 0.01$[b] |
| 4 | $p < 0.01$[a] | |
| 5 | $p < 0.01$[a] | |
| 6 | $p < 0.01$[a] | |
| 7 | $p = 0.04061$[b] | |
| 8 | $p = 0.005843$ | |
| 9 | $p = 0.04314$[b] | |
| 10 | $p = 0.02301$[b] | |

[a] a highly significant value ($p < 0.01$).
[b] Indicates a significant value ($p < 0.05$).

AQ 5




AQ 6 **Figure 8.6 Pairwise Pearson correlationsamong variables for the data-sets. Circle size is proportional to correlation coefficient *r*. Circle color indicates significance: gray, nonsignificant; blue, $p < 0.05$; green, $p < 0.001$; red, $p < 0.0001$ (a) Data-set 1 and (b) Data-set 2.**

## *Methodology for Experiments*

From the data-sets, specific samples were hidden to simulate missing values in three different fashions:

- Random Missing Data (RMD): Ghosted samples were chosen at random. Ghosted data accounts for 10% of each variable.
- Random Missing Blocks (RMB): Ghosted samples were chosen in blocks to have consecutive subseries of missing data. Ghosted data accounts for 10% of each variable. However, the location of the ghosted block and the number of blocks is random.
- All Missing Data (AMD): One full variable was ghosted at a time. Reconstruction can only occur from related information.
- For each fashion, 10 train/test pairs were prepared for a 10-fold cross-validation. Note that the AMD has *d* test for each train case where *d* corresponds to the number of variables in the data-set. After preparation of the ghosted test data-sets, reconstruction was attempted by means of the following techniques:
- Static Bayesian Network (BN). Discretization was set to 5 equidistant intervals. Structure was learned using the PC algorithm (Spirtes et al., 2000).
- Autoregressive Bayesian Network (AR-BN). Autoregression order was fixed to $<p, q>=<1,1>$. Vertical (intra-stage) structure was learned using the PC algorithm.

Equidistant intervals were used at all times, with the number of intervals being either 4 or 5, as bounded by memory limitations. The exemplary network for the data-set 2 is illustrated in Figure 8.5.

- Linear interpolation (LI).
- Cubic spline interpolation (CSI).
- Autoregressive Models (AR(1)).
- Autoregressive Models (AR(*n*)). Order *n* was chosen according to Table 8.1. Notwithstanding, during the preparation of the train/test sets, some of the test sets did contain a number of samples lower than the autoregressive order, i.e. AR order 25 for data-set 1, variables 1 and 3. In those cases, the highest possible order was chosen based on the number of available samples.

As indicated earlier, for each reconstruction technique and ghosting fashion, a 10-fold validation was made. Since the AMD scenario can only be reconstructed from related information, this scenario cannot be resolved by interpolation or autoregressive models.

Therefore, the experiments were applied in two data-sets, in three scenarios, using six techniques, and repeated 10 times. In total, 280 simulations were executed using MATLAB and Hugin (Andersen et al., 1989). For 3 simulations, mistakes in the pipeline from training to test were detected, and their results not included for further analysis. Statistical analysis was carried out in R[3].

## *Results and Discussion*

An example of the reconstruction with the different techniques is illustrated in Figure 8.7. The red line indicates the original time-series with the complete values. The rest of the plots correspond to all other techniques: static Bayesian network, AR-BN, linear interpolation, spline interpolation, AR(1), and AR(*n*). The three graphs correspond to the reconstruction of the three variables of data-set 2. The experiments correspond to the RMD scenario, that is, every element of the three time-series of data-set 2 is considered missing and is estimated with all the techniques covered in this chapter. The evaluation is made with respect to the observed element (red line). The solid line corresponds to AR-BN technique. Qualitatively, AR-BN performs well, especially for variables 1 and 2. Linear interpolation also has a nice performance. The difference between linear interpolation and AR-BN is that the first performs well when the variable has low correlation between each other, while AR-BN takes into account the relation between all variables in a domain.

Figure 8.8 summarizes the errors incurred by each technique, according to the error metrics described in the section "Error Metrics for the Estimation of Missing Values." Bars correspond to average values and error lines indicate standard deviation.

[3] R is a language and environment for statistical computing and graphics, see http://www.r-project.org/.

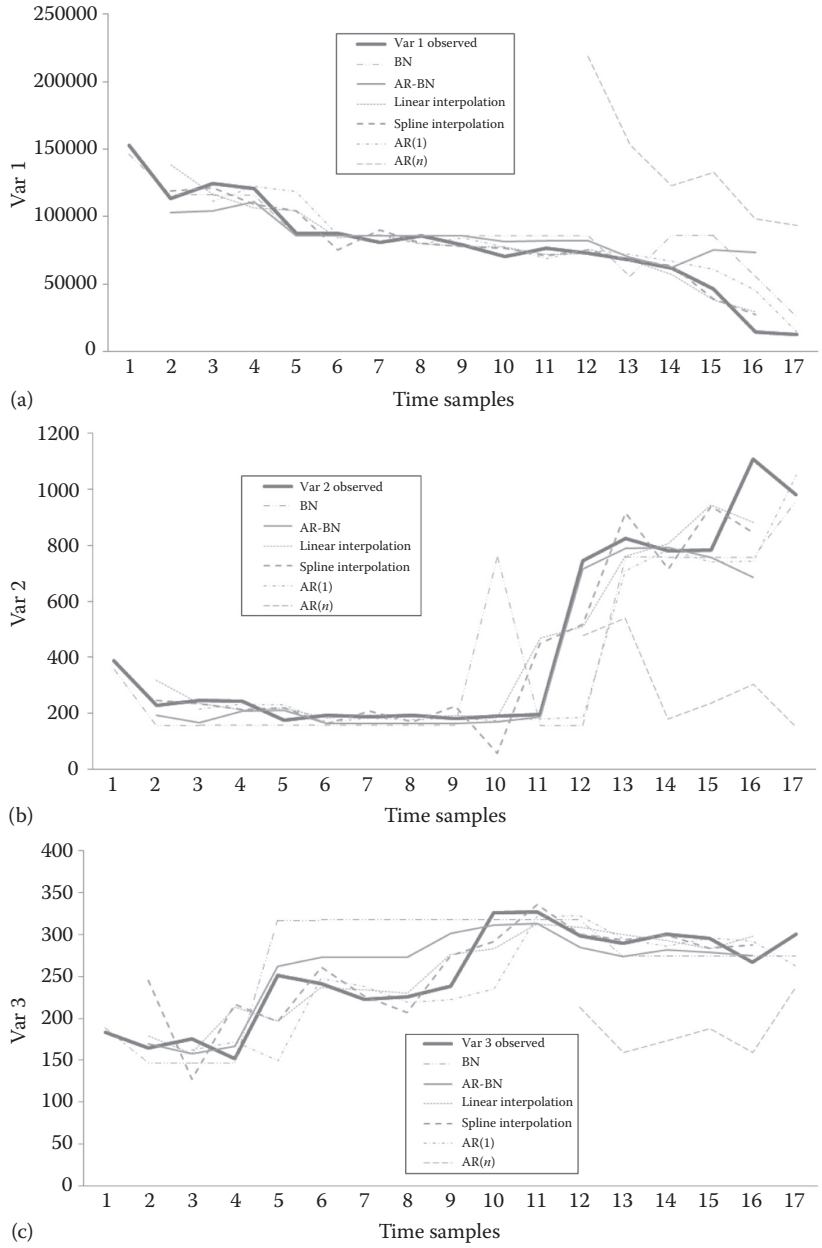



**Figure 8.7 Example of the missing data estimation using the different techniques. The example corresponds to the 3 different variables of data-set 2, respectively, for an RMD scenario. For this example, each sample of the time-series is hidden one at a time, and the missing sample is estimated using the rest of the series as necessary by the different estimation techniques.**

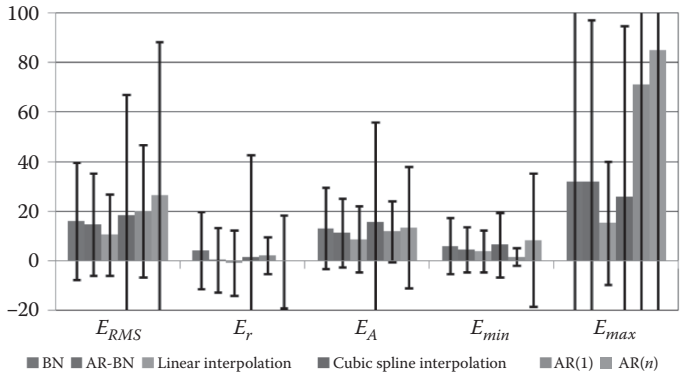



**Figure 8.8 Reconstruction errors incurred by each technique across data-sets, scenarios, folds, and variables. Bars and error lines correspond to average values and standard deviation, respectively.**

Figure 8.9 provides a more detailed view by data-set and error metric. The left column in the figure corresponds to data-set 1, and the right column corresponds to data-set 2. The rows in the figure correspond to every error metric described in the section "Error Metrics for the Estimation of Missing Values." First, $E_{rms}$, next $E_r$, $E_a$, $E_{min}$, and $E_{max}$. Bars and error lines correspond to average values and standard deviation, respectively. Inside each graph, the three scenarios of missing values are exposed: random missing data (RMD), random missing blocks (RMB) and all missing data (AMD), as explained earlier.

From this detailed view, it can be appreciated that the proposed AR-BN achieves a good compromise in the reconstruction across different scenarios, data-sets, and error metrics. Unexpectedly, linear interpolation achieves better overall reconstruction than the more advanced spline interpolation. Classical autoregressive models achieve reasonable performance but are highly unstable in their predictions as demonstrated by the large standard deviations coupled with disparate differences between $E_{min}$ and $E_{max}$.

In order to clearly understand the meaning of these results, let us revise a portion of the information in Figure 8.9. Consider the left column, corresponding to data-set 1 formed by 10 variables. The first three rows show the measured performance of all methods with three parameters: root mean square, average percent relative, and absolute average percent relative. Literature considers that the absolute percent relative error ($E_r$) is an important indicator of the accuracy of the models (Osman et al., 2001). In this indicator (second row of Figure 8.9), AR-BN obtains low values for missing data and missing blocks compared with other traditional methods. However, for comparing the absolute average percent relative error ($E_a$), the values favored other methods.

Considering the parameters ($E_min$) and ($E_max$), the minimum values are better for methods that have no interplay between variables, but the $E_max$ parameter

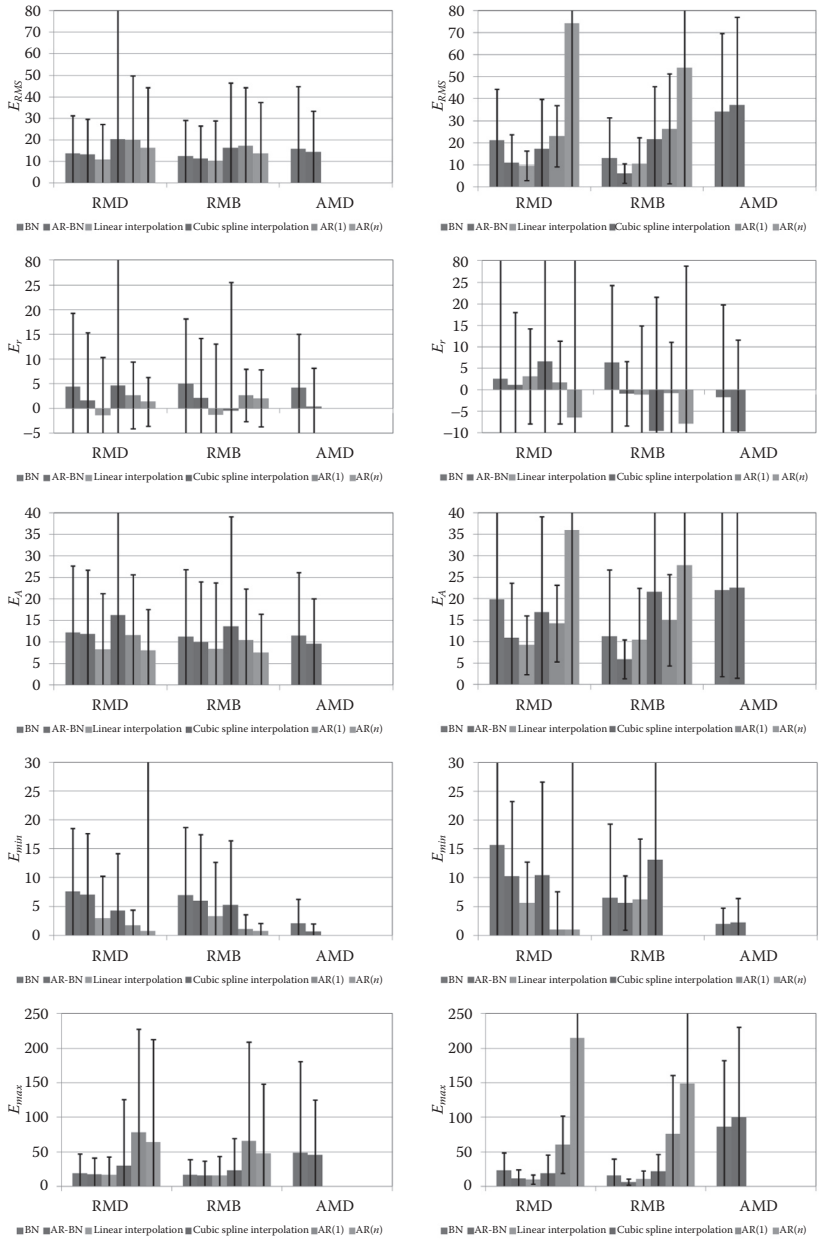



**Figure 8.9 Reconstruction errors incurred by each technique across folds and variables. Columns correspond to data-set; Left: data-set 1; Right: data-set 2; Rows correspond to different error metric: from top to bottom: $E_{rms}$, $E_r$, $E_a$, $E_{min}$, and $E_{max}$. Bars and error lines correspond to average values and standard deviation, respectively.**

favored methods that considered relations between the variables in the data-set. Notice the bottom-left graph of Figure 8.9; the performance of all these methods are similar for both missing data and missing block.

## *Limits of Missing Data Estimation Approaches*

Figure 8.10 relates the variable feature space given by the variable autoregressive order and its average relation to all other variables in its data-set (avg_r) against the

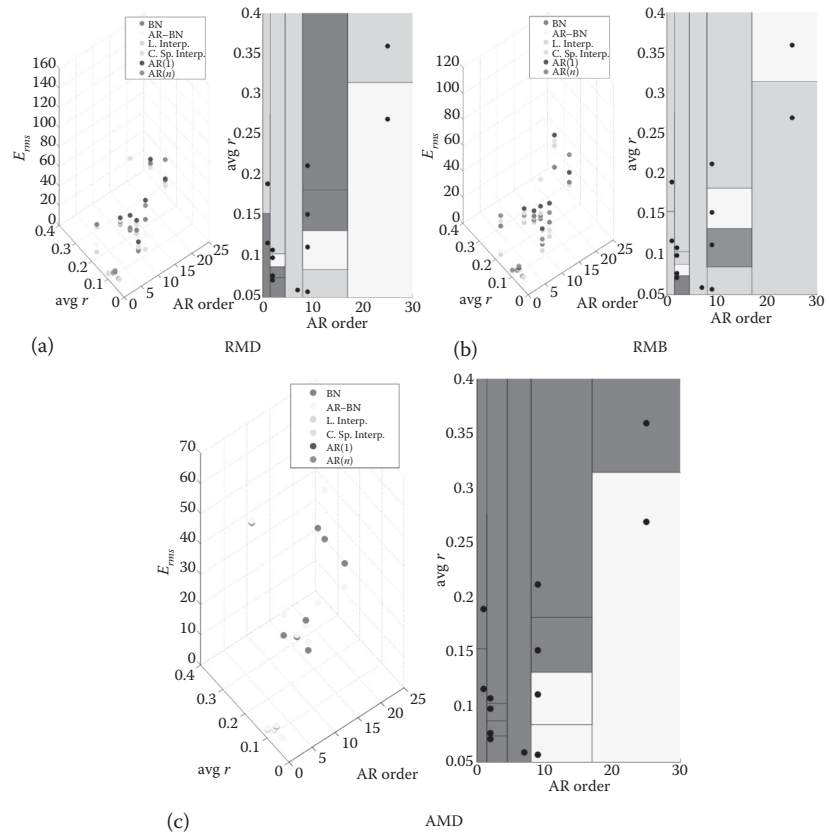



**Figure 8.10 Relation between the variable feature space and the techniques for the three different scenarios. Left: Scatterplots of the variable feature space versus the error for each variable reconstructed through different techniques. The technique that achieves the lowest error is considered to dominate the region of the variable feature space. In order to determine the region of the variable feature space, the different feature vectors for each of the variables for both data-sets are used as seed-vector quantizers for establishing a Voronoi partition. Each region of the Voronoi parcellation is then colored according to the dominant technique.**

AQ 7

AQ 8

dominant technique for the three scenarios: (1) for missing data, (2) for missing block, and (3) for missing variable. On the left side of each scenario is a scatter plot of the variable feature space versus the error for each variable reconstructed through different techniques. The technique that achieves the lowest error is considered to dominate the region of the variable feature space. The dominant technique is that which affords the lowest error in a particular region of the variable feature space. On the right side, regions are calculated using the Voronoi partition. In order to determine the region of the variable feature space, the different feature vectors for each of the variables for both data-sets are used as seed-vector quantizers for establishing a Voronoi partition. 10 seed vectors appear from data-set 1 plus 3 seed vectors for data-set 2. Each region of the Voronoi parcellation is then colored according to the dominant technique.

It can be appreciated how the use of one technique over the other is subjected to the characteristics of the variable in terms of its autoregressive information, as well as the amount of dependency that the variable shares with fellow variables in the data-set as hypothesized. In particular, linear interpolation performs particularly well when the estimated autoregressive orders of the variables are low. When a full variable needs to be reconstructed from related information (scenario (c)), it is obvious that the AR-BN dominance of the variable feature space grows as the autoregressive information does.

## Conclusions and Future Work

We have explored the relation between a variable feature space represented by its autoregressive order and its relation to other variables in its data-set against different reconstruction techniques. Our results suggest that the interplay between the variable's characteristics in the data-set dictates the most beneficial reconstruction option.

We have shown that the proposed AR-BN achieves a particularly competitive reconstruction regardless of the scenario, data-set, and error metric used. Although we have reported signals stationarity for reproducibility, it has not further been considered for this chapter. We believe signal stationarity will also be a critical element in the variable feature space, supporting the decision of which estimation technique to use. Consequently, we plan to explore its effect.

The AR-BN model can be trivially extended to any level of autoregression and can be easily adapted for nonnumerical data. In this sense, different autoregressive stages, whether past or future, must be added "in parallel" rather than "in series" so that these observations can be appreciated through the Markov blanket. We believe the proposed AR-BN profits from both within-variable information and statistical dependencies across variables, thus representing a valuable tool for the estimation of missing data in incomplete databases.

## References


Abraham, B., and G. E. P. Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 66(2):229–236, 1979.

Akaike, H. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247, 1969.

Andersen, S. K., K. G. Olesen, F. V. Jensen, and F. Jensen. Hugin: A shell for building bayesian belief universes for expert systems. In *Proceedings of the Eleventh Joint Conference on Artificial Intelligence, IJCAI*, pp. 1080–1085, Detroit, MI, August 20–25, 1989.

Balke, N. S. Detecting level shifts in time series. *Journal of Business & Economic Statistics*, 11(1):81–92, 1993.

Caruso, C. Can a social-media algorithm predict a terror attack? *MIT Technology Review*, June 16, 2016.

Chatfield, C. *The Analysis of Time Series: An Introduction*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

de Boor, C. *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer-Verlag, New York, 1978.

Dean, T., and K. Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3):142–150, 1989.

Dempster, A. P., N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B* (*Methodological*), 39(1):1–38, 1977.

Fukunaga, K., and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20(2):176–183, 1971.

Hernández-Leal, P., L. E. Sucar, and J. A. González. Learning temporal nodes bayesian networks. In *Twenty-Fourth International Florida Artificial Intelligence Research Society Conference* (*FLAIRS'2011*), pp. 608–613, Palm Beach, FL, May 18–20, 2011. Association for the Advancement of Artificial Intelligence.

Herrera-Vega, J., F. Orihuela-Espina, P. H. Ibargüengoytia, E. F. Morales, and L. E. Sucar. On the use of probabilistic graphical models for data validation and selected application in the steel industry. *International Journal of Approximate Reasoning, In revision.*, 2012.

Hoo, K. A., K. J. Tvarlapati, M. J. Piovoso, and R. Hajare. A method of robust multivariate outlier replacement. *Computers and Chemical Engineering*, 26:17–39, 2002.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

Ibargüengoytia, P. H., M. A. Delgadillo, U. A. García, and A. Reyes. Viscosity virtual sensor to control combustion in fossil fuel power plants. *Engineering Applications of Artificial Intelligence*, 29:2153–2163, 2013a.

Ibargüengoytia, P. H., U. García, F. Orihuela-Espina, J. Herrera-Vega, L. E. Sucar, E. F. Morales, and P. Hernández-Leal. On the estimation of missing data in incomplete databases: Autoregressive bayesian networks. In *The Eighth International Conference on Systems, ICONS-2014*, Sevilla, Spain, 2013b. IARIA.

Ibargüengoytia, P. H., S. Vadera, and L. E. Sucar. A probabilistic model for information and sensor validation. *The Computer Journal*, 49(1):113–126, 2006.

Koller, D., and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.


AQ 9

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54:159–178, 1992.

Lamrini, B., E.-K. Lakhal, M.-V. Le Lann, and L. Wehenkel. Data validation and missing data reconstruction using self-organizing map for water treatment. *Neural Computing and Applications*, 20:575–588, 2011.

Lancaster, P., and K. Salkauskas. *Curve and Surface Fitting: An Introduction*. Academic Press, London, UK, 1986.

Marr, D., and Hildreth, E. Theory of edge detection. *Proceedings of the Royal Society London B*, 207:187–217, 1980.

Mihajlovic, V., and M. Petkovic. Dynamic bayesian networks: A state of the art. CTIT technical report series TR-CTI 36632, University of Twente. Department of Electrical Engineering, Mathematics and Computer Science (EEMCS), 2001.

Muirhead, C. R. Distinguishing outlier types in time series. *Journal of the Royal Statistical Society. Series B* (*Methodological*), 48(1):39–47, 1986.

Osman, E. A., O. A. Abdel-Wahhab, and M. A. Al-Marhoun. Prediction of oil pvt properties using neural networks. In *SPE Middle East Oil Show*, 14 pp., Manama, Bahrain, March 17–20, 2001. Society of Petroleum Engineers (SPE).

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, 1988.

Peng, J., S. Peng, and Y. Hu. Partial least squares and random sample consensus in outlier detection. *Analytica Chimica Acta*, 719:24–29, 2012.

Sato, H., N. Tanaka, M. Uchida, Y. Hirabayashi, M. Kanai, T. Ashida, I. Konishi, and A. Maki. Wavelet analysis for detecting body movement artifacts in optical topography signals. *NeuroImage*, 33:580–587, 2006.

Shibata, R. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63:117–126, 1976.

Spirtes, P., C. Glymour, and R. Sheines. *Causation, Prediction and Search*. MIT Press, Cambridge, MA, 2000.

Sucar, L. E. *Probabilistic Graphical Models: Principles and Applications*. Advances in Computer Vision and Pattern Recognition. Springer, London, UK, 2015.

Tsay, R. S. Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, 7:1–20, 1988.

Tuntun, S., M. T. Khasawneh, and J. Zhuang. New framework that uses patterns and relations to understand terrorist behaviors. *Expert Systems with Applications*, 78:358–375, 2017.

Vagin, V., and M. Fomina. Problem of knowledge discovery in noisy databases. *International Journal Machine Learning & Cyber*, 2:135–145, 2011.

Walczak, B. Outlier detection in multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 28:259–272, 1995.

## Author Query Sheet

**Chapter No.: 8**

| Query No. | Queries | Response |
|---|---|---|
| AQ 1 | Please confirm whether the shortened running head is okay. | |
| AQ 2 | Please confirm whether "signal own information" can be set as "signal-owned information"? | |
| AQ 3 | Please confirm Hipothesis can be changed to "Hypothesis". | |
| AQ 4 | Please confirm "complete train set" is fine in the sentence "During the learning process, a complete…" | |
| AQ 5 | Please confirm the fixed table footnotes are fine. | |
| AQ 6 | As the figure is printed in gray scale kindly provide the caption without mentioning of color in Figure 8.6. | |
| AQ 7 | Please provide part labels for Figure 8.10. | |
| AQ 8 | As the figure is printed in gray scale kindly provide the caption without mentioning of color in Figure 8.10. | |
| AQ 9 | Please update year of publication, volume number and page number for the reference "Herrera-Vega et al. (2012)." | |