Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 2 (2014) pp. 128-139

Forecasting in Port Logistics and Economics using Time Series Data Mining Model

Ana Ximena Halabi Echeverry[1,2], Deborah Richards[1], Ayse Bilgin[4] and Jairo R. Montoya-Torres[2]

[1] Computing Department, Macquarie University,
Sydney NSW 2109, Australia
(ana.halabiecheverry, deborah.richards) @mq.edu.au

[2] Escuela Internacional de Ciencias Economicas y Administrativas, Universidad de La Sabana,
km 7 autopista norte de Bogotá, D.C., Chía (Cundinamarca), Colombia
(ana.halabi, jairo.montoya) @unisabana.edu.co

[3] Statistics Department, Macquarie University,
Sydney NSW 2109, Australia
ayse.bilgin@mq.edu.au

***Abstract*: This paper addresses the question of how to develop forecasting models resulting from business processes that can be embodied in an intelligent decision support system. Moreover the design is suitable for evolving logistics and economic situations in which ports plan or foresee to have an improved economic role. The work presented in this paper also forms one component of a conceptual intelligent decision-making support system (*i*-DMSS) for port integration. The key objective of this work is to offer a model-based approach to Time Series Data Mining (TSDM) based on the assumptions that the time series may be produced by an underlying model, and that its flexibility is suitable to perform multivariate time-series analysis encompassing the notion of model selection and statistical learning known as the core of forecasting systems. The interrelated activities to induce domain knowledge are specified as the data collection principles, the descriptive modelling and normative modelling. Results indicate that for the period 2001 to 2005, the commodity throughput of coffee (tons) handled in the port of Buenaventura gains importance in the prediction of the Colombian national exports of coffee, thus indicating that the port operation was able to affect the economy in this regard. The previous period was strongly affected by outliers, creating a random walk process difficult to fit but feasible to produce due to unstable conditions evidenced in the economy.**

***Keywords*:** Time Series Data Mining, Forecasts, Port Economics, Port Logistics, Buenaventura Port**.**

I. Introduction

To deal with the complexity of port operations higher resolution and forecasting models using time series data are paramount and necessary. In this paper we pose the question of how to develop forecasting models resulting from business processes that can be embodied in a computerized system or decision support system (DSS) for port integration (i.e, concept to understand harmonised/standardised strategic levels of information among several ports). Moreover we address the question of how to design forecasting models of evolving logistics and economic situations in which the port can play an improved role. The development of DSS so far, "have focused on high-level decision making (strategic decision) but using low levels of representation (data, equation, etc.) because (1) the notion of data representation has not been sufficiently studied and (2) high level decisions are more appealing than small decisions" [27, p.27]. The necessity of introducing more intelligent decision support tools to meet future challenges of ports becomes evident from Klodzinski and Al-Deek [1, p.2] who present ports as "important economic generators" [for whom] infrastructure improvements [physical and informational] is essential to accommodate potential growth". Tiwari [36] also remarks that efficiency is essential for ports to become players in the international arena since ports are the trunk of any country's foreign trade. Tiwari's framework shows the close relationship between port performance and its environment and notes that any changes in operations, even small ones, would impact the larger national economy.

A primary approach found in the literature that seeks to fill the gap between port performance and its environment, i.e., market and local economy, is the use of statistical learning based on data and hypotheses [23]. However, the concept of learning under uncertain conditions from observations brings challenges in time-dependent domains because unexpected relationships and patterns need to be learnt [4, 17]. Many of the learning tasks involve prediction as stated by Batyrshin and Sheremetov [3]. This paper innovatively introduces the concepts of 1) learning under time and uncertainty and 2) learning from incomplete data, as two general frameworks traditionally not defined as computational learning among the statistical methods in literature of Artificial Intelligence (AI).We propose these frameworks as a means for improved technological analyses of issues concerning lack of investment facing ports such as Buenaventura Seaport (BUN) in Colombia, whose operations should improve with broader economic analyses. Finding ways to discover critical information under uncertain scenarios, may better inform decision-making of ports, especially those located in

developing countries.

Literature suggests that parametric modelling approaches may not provide a suitable representation for port managerial decisions [24]. However, this criticism is largely based on the tendency for non-parametric models used in decision-making to use expert systems approaches which raise concerns with handling of uncertainty [5]. As an alternative to the expert-system (i.e. rule based) approach, this paper develops a model-based approach to handle time series data for seaports. The model we develop allows the use of time series data mining for knowledge discovery and decision making. "Adaptive and innovative application of the principles and techniques of classic data mining in the analysis of time series resulted in the concept called time series data mining (TSDM) [25, p.7]". Previous work of Povinelli [28] situates the TSDM concept in between time series analysis, chaos and nonlinear dynamics, data mining and genetic algorithms. He states that "from time series analysis comes the theory for analyzing linear, stationary time series...from dynamical system comes the theoretical justification of the TSDM methods…from data mining comes the focus on discovering hidden relationships and patterns, [and] from genetic algorithms a robust and easily applied optimization method [28, p.10]".

The fundamental problem in TSDM is how to represent time series data. In a recent review of TSDM methods, Fu [10] mentions that four mining tasks can be classified in time series representation, they are: pattern discovery and clustering, classification, rule discovery and summarization. Wang et al. [38] propose that pattern discovery in time series with high dimensionality can be performed by a feature extraction process. Among these features trend, seasonality, skewness, kurtosis and chaos are of primary interest in this paper. Moreover, Ratanamahatana et al. [29] and Fu [10] remark that TSDM must be able to deal successfully with dimensionalities in hundreds or thousands. That might be why there is a lack of decision support systems for multivariate data series that can be used in different problem domains [3, 18]. Consequently, the key objective of this work is to offer a model-based approach to TSDM based on the assumption posed by Esling & Agon [9] that the time series may be produced by an underlying model. These authors point out that several parametric temporal models are available in TSDM including statistical modelling by feature extraction, ARMA models and Markov Chains (MCs). We developed an ARIMA model (ARIMA stands for multivariate auto-regressive, integrated, moving average) to find parameters of such a model representation.

The following sections of this paper have been organised as follows. Section II provides some background to the domain, its disciplinary characteristics and forecasting methods in use. The research approach and the dataset are introduced in Section III. In Section IV, we present our knowledge discovery approach based on searching for critical information through the analysis of hidden patterns in the series. Section V outlines the phases and modules required to generate the data mining tasks and empirical analyses introduced and follows a more aggregated structure called the conceptual intelligent decision-making support system (*i*-DMSS) for port integration. Section VI presents conclusions.

II. Background and Related Literature

A. Disciplinary Characteristics

In this paper we seek to model historical transactions such as port throughput and foreign trade by commodity for the purpose of identifying specific ports playing an important or noteworthy role in port economic sustainability and operational efficiency. The choice of key port operating variables to denote impacts on the larger national economy are relative to those under the control of port management and its relationship with port performance (see Figure 1).

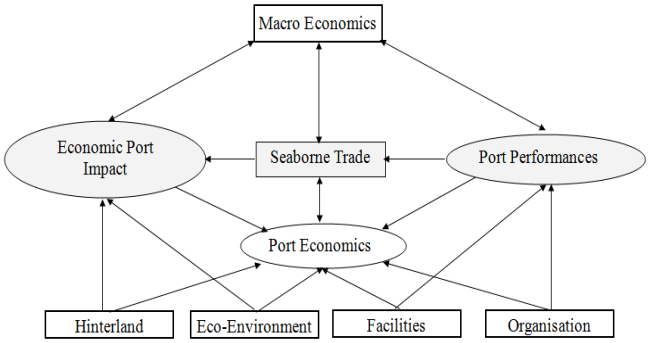

Figure 1. Relationship between Port Performances and Port Economic Impacts. Adapted from Tiwari et al. [34, p.2]

De Langen et al [8] mentions that important explanatory variables are associated with container terminals and throughput measures whilst Talley [32, pp.508-509] details a number of port performance indicators (PPIs) with respect to the maximisation of the annual throughput in a competitive context, as follows:

1. Annual average port charge per throughput ton (for a given type of cargo).
2. Annual average ship loading service rate (for a given type of cargo), i.e. tons of cargo loaded on ships per hour of loading time.
3. Annual average ship unloading service rate (for a given type of cargo), i.e. tons of cargo unloaded from ships per hour of unloading time.
4. Annual average loading service rate (for a given type of cargo) for port vehicles of inland carriers, i.e. tons of cargo loaded on port vehicles per hour of loading time.
5. Annual average unloading service rate (for a given type of cargo) for port vehicles of inland carriers, i.e. tons of cargo loaded on port vehicles per hour of unloading time.
6. Annual average daily percent of time that the port's channel adheres to authorised depth and width dimensions (port channel accessibility indicator).
7. Annual average daily percent of time that the port's berth adheres to authorised depth and width dimensions (port accessibility indicator).
8. Annual average daily percent of time that the port's channel is open to navigation (port channel reliability indicator).
9. Annual average daily percent of time that the port's berth is open to the berthing of ships (port berth reliability indicator).

10. Annual average daily percent of time that the port's entrance gate is open to inland-carrier vehicles (entrance gate reliability indicator).
11. Annual average daily percent of time that the port's departure gate is open to inland-carrier vehicles (departure gate reliability indicator).
12. Annual expected probability of damage to ships property while in port.
13. Annual expected probability of loss of ships property while in port.
14. Annual expected probability of damage to inland-carrier vehicles while in port.
15. Annual expected probability of loss to the property of inland-carrier vehicles while in port.
16. Annual expected probability of damage to shippers' cargo while in port.
17. Annual expected probability of the loss of shippers' cargo while in port.

A basic economic model of the port is presented by Talley [32, p. 45]. He specifically points to an economic theory of the cargo port and the applicability of the theory to port performance. A decade ago, Talley [34] argued that if a competitive environment exists among cargo ports, they should consider their economic throughput in evaluating performance. Assuming that the cargo basically handled by the port is bulk and container cargo, with bulk referring to dry and liquid cargo (i.e., liquid chemicals), and container cargo referring to stored standardised cargo boxes (generally of 20 and 40 feet), N_b measures the annual bulk throughput and N_c measures the annual container throughput at the port. The basic model represents the annual demand of bulks and container throughputs as functions of their generalised prices, where GP_b is the sum of the port's prices per unit of throughput of bulk cargo and GP_c represents the sum of port prices per unit of throughput of container cargo. These calculations are presented in Equations (1) and (2):

$$N_b = N_b(GP_b) \quad (1)$$
$$N_c = N_c(GP_c) \quad (2)$$

An advantage to port management in controlling port throughput is that it can be evaluated over time. Talley [33] mentions that the same occurs with any other performance indicator as other services will move the port to achieving an economic position. Among the research gaps found in the literature, the throughput volume seems to point out current limitations in deducing port performance in relation to economic objectives [7]. However this paper acknowledges that the annual throughput volume is a performance standard, thus integrating mechanisms among ports must consider an independent analysis of this variable and suggests that this approach helps to assess multi-port performances for a given time period. In Australia for example, a number of PPIs have been selected to measure economic reforms by government. Tull and Affleck [37] point out that in Western Australia (WA) port authorities assess their performance, functions and economic impacts for competitive purposes. They also suggest that "there is no exclusive link between economic regulation and superior physical and economic performance [of the port] (p.16)". This quotation indicates the necessity of linking business practices with economic factors giving a wider perspective of the port. Similarly, Tiwari [36, p.19] delineates ten aspects to foster economic development and port sustainability, three of these aspects are: 1) have a proper vision for growth, 2) plan for the long-term, and 3) have commitment and show political will.

For this paper, it is assumed that the specific choice of variables is induced for a port with the economic objective of increasing throughput while positively affecting the local economy, and thus the very first aspect on Talley's list (i.e., annual average port charge per throughput ton (for a given type of cargo) is selected to illustrate the use of this PPI. However the major difference with Talley's main purpose is that, the indicator-selection is used to measure the impact of the port's operation in the local economy instead of addressing the use of PPIs to evaluate economic objectives of the port such as profits and deficits.

B. Review of Port Forecasting Methods

Forecasting has been a common practice to improve seaport competitiveness. For example Hong Kong, nowadays one of the busiest container ports in the world, has been using a regression analysis approach to forecast port general cargo throughput (a sum of cargo discharges and loadings in gross weight of tons handled by the port) for its port planning and development over the decades [21]. Very few approaches offer solutions to relate the efficiency of operations with the economic impact analysis of the ports. For example, Lam et. al. [21] created and used two datasets: one to forecast the freight movement and the other to bring together explanatory factors associated with commodities and shipment type. Their study used neural networks to create predictions; however, some discrepancies of predictions were associated with incorrect econometric projections of the explanatory factors being no longer significant due to changes in the economy. They performed a reliability analysis using Monte Carlo simulation to correct the prediction errors associated with forecasted freight movements. In a second example, Zondag et. al. [39] developed a forecasting approach that simulates port competition. In their model, traffic of freight cargo is the explanatory variable under scenarios of port infrastructure. They use the multinominal logit model incorporating uncertainties about the factors that determine route choice by shipping companies. The shortcoming of this model is that it is based on discrete choices therefore the authors see the necessity to extend the modelling approach.

III. RESEARCH APPROACH

In this paper we are interested to specifically answer the question: Which data-driven mechanisms can be implemented to identify ports playing an improved logistic role in local economies? We seek to model the business intelligence processes relevant to answering this question and seek to distinguish between what decisions need to be made for port integration and the procedures that are needed for the modelling process. According to the classical approach of Stabell [31], processes on the decision-making side of the DSS rely on three fundamentals: (1) data collection principles, (2) descriptive modelling and (3) normative modelling, these three fundamentals are provided in this section. The

remaining subsections further develop the model-based approach and application of TSDM.

The work presented in this paper also forms one component or module of a conceptual intelligent decision-making support system (*i*-DMSS) for port integration. The complete schema of the *i*-DMSS for port integration involves three phases. The first phase commences with extraction of data from 13 heterogeneous repositories. The second phase involves the creation of customised datasets, each one tailored to the specific decision on port integration to be explored. Depending on the nature of the decision, some data sets do not include a time dimension but only look at the ports at a certain point in time without considering the impact of time on the relationships – similar to cross-sectional designs for medical data. For the question explored in this paper we have used a dataset that assumes a time-dependent relationship and records historical data as series data. The third phase of our port integration schema involves the implementation of data mining algorithms to discover patterns through classification(s) and prediction(s). Each data mining task is performed with the help of statistical software, namely: IBM SPSS version 19.0, Rapid Miner version 5.0 and R version 2.13.1/Amelia II. The final phase considers the interoperability between data-driven models using predictive modelling markup language (PMML) as will be further explained later in this section under Interoperability.

The concept of an *i*-DMSS module for port integration is associated with a data mining algorithm on which the model is trained. It also accomplishes specific business functions for the port integration that can be regarded as business intelligence (BI) tools. An *i*-DMSS module applies a particular data mining task to provide information for accomplishing the decision of port integration [18]. Modules are designed on a particular dataset and for a specific purpose. Each module operates on different subsets of data, which provides opportunity to show the scope of this kind of DSS. Module III encompasses four series spanned initially from a long segment covering the years 1999 to 2012. The segments selected are 1999 to 2003 and 2001 to 2005 and the series describe: CoffExpCol (national exports of coffee) and CoffOutBUN (port throughput in tons for coffee cargo). Each segment is treated as a case in this module.

Two linked activities are part of the approach to automatically induce domain knowledge: 1) completion of datasets through multiple imputation (MI) of missing values, and 2) development of a forecasting model through Time Series ARIMA. Table 1 lists the two linked activities or data-driven models to induce domain knowledge.

Linked activities to induce KDD	DM Software
Multiple Imputation of Missing Values	R, Amelia II
Time Series Forecasting ARIMA	IBM SPSS 19

Table 1. Linked activities (data-driven models) to induce domain knowledge in Module III

MI is an important pre-processing step that also can be regarded as a learning procedure able to produce unbiased estimates and a suitable level of uncertainty in the model [22]. MI is also a sound statistical approach in which multiple samples are drawn from the population and taken to make inferences over missing values. Ni et al. [26] state that MI is able to learn simultaneously from correlations and estimates within and among variables. That is why it is known as an unbiased estimation method. By this, an understanding of how much confidence can be placed in the imputation result is possible.

Subsection C introduces an exhaustive method of preprocessing and imputation taking into account appropriate assumptions when working with missing values in series data. We introduce these analyses following authors such as Honaker and King [15] that agree in pointing several issues and challenges on data drawn from developing countries such as incompleteness due to high cost of collection.

The second activity (development of a forecasting model through Time Series ARIMA) is focused on discovering hidden relationships and patterns from series data. Time Series Data Mining (TSDM) encompasses the notion of model selection and statistical learning that take a novel approach to the understanding of outliers and the identification of rules and patterns that characterise the time series such as trend, seasonal, cyclical variations and random moves [8, 29].

A. Data Collection Design and Analysis

We tested the framework suggested by Zongag et al. [39] by, collecting data from the South American Port of Buenaventura as a case study. Zongag et al's main point is to offer a new port forecasting approach that models port competition by considering changes in the market such as impacts on policy measures, infrastructure projects and pricing. They distinguish the components that can be followed in Figure 2: scenario/policy settings, the data traffic flows, the trade growth model, and the route/ logistic chain choice model. Each component uses mathematical modelling with the exception of a trade growth model that uses agent-based simulation modelling. This component simulates one year at a time, starting in 2006 with the capacity of continuing indefinitely and uses disaggregated trade flows in tons between country pairs. The current available framework does not encompass the notion of model selection and statistical learning known as the core of forecasting systems [22]; consequently Zongag et al's [39] study is an example of forecasting models using low levels of representation such as data and equations.

To apply the port competition forecasting framework in Figure 2, we consider the relationship between the components: the data traffic flows within the port (port throughput) and the trade growth model by commodity. The port throughput in tons (for coffee cargo) is used as the internal factor (in control) by the Buenaventura (BUN) port, while the foreign trade of coffee or the national exports of this commodity (Colombia) is used as the external factor that may be influenced in its relation with the port throughput. In the majority of cases analyses often produce forecasts by commodity under the assumption that the port performance does not influence trade flows [20]. In our approach, the forecasting analysis produces influences from the port to the local economy, in order to raise confidence within partner

ports about trade facilitation and port performance in the context of the decision of port economic sustainability and operational efficiency for port integration.

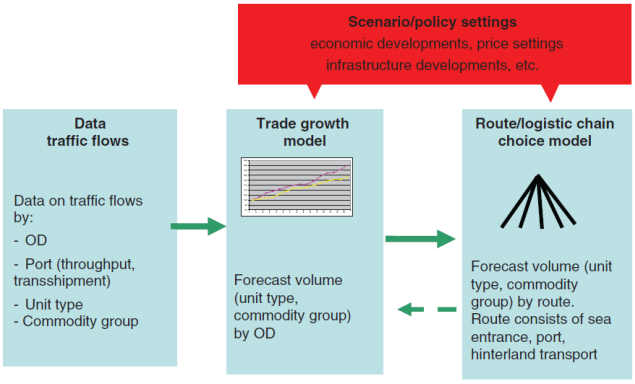

Figure 2. Port competition forecasting framework [37]

We introduce the main steps of analysis and modelling using a customized dataset for BUN Port. Initially data was collected between 1999 and 2012; however taking into consideration the large variability of changes in the economy over that time period, such a prolonged series was not useful. Thus, we split the series into equivalent cross-segments of five years each to provide sufficient information and allow us to make comparisons. Analyses presented in this paper correspond to the first two segments, i.e. 1999 to 2003 series and 2001 and 2005 series. In accordance with De Langen et al. [7] who provide guidance for port decision-making, important explanatory factors associated with container terminals are throughput measures [9]. Thus, in order to predict the Colombian coffee national exports, we have made the appropriate empirical observations over 120 months in relation to the port's throughput for this commodity. Data was taken from a local port repository (www.sprbun.com) and the system of the Department of Customs in Colombia (DIAN) (http://websiex.dian.gov.co)

B. Descriptive Modelling

The objective of the descriptive model is to identify the components forming the informational structure for the decision of port economic sustainability and operational efficiency in the context of port integration. In this paper the dimension of analysis is the logistics operation at a local port; however the context of analysis is either local or country based on its relationship with the foreign trade of coffee (the national exports of this commodity). In this module it is possible to induce further study of a specific port in which the desire of integration leads to an understanding of the economic impact of such a port due to its operational efficiency and consequently leads to a better choice of partners.

C. Normative Modelling: Time series Data Mining (TSDM)

In this subsection we will introduce a step-by-step TSDM method. According to Milanovic et al. [25], the importance of TSDM is related to emergent phenomena in different research areas, including those in business and economics. A number of applications predate TSDM tasks such as classification, indexing (query by content), clustering, prediction (forecasting), summarising, anomaly detection and segmentation. All aspects of TSDM are centred on the identification of rules and patterns that characterise the time series such as trend, seasonal, cyclical variations, and random moves. The importance of TSDM stems from its flexibility to perform multivariate time-series analyses. Esling and Agon [8] point out that TSDM encompasses the notion of model selection and statistical learning known as the core of forecasting systems. This robust approach overcomes issues such as the uniqueness of each record in a time-series. We make novel use of TSDM distinctiveness by understanding and retaining outliers in series data, particularly if they contain essential information for decision making and imply a loss of understanding to improve decisions; also, by analysing patterns and predictors' impacts in the series. TSDM offers the possibility to get better insights into the trends by partitioning the series into subsequent small series in a repetitive process that helps to capture patterns whilst learning the modelling task. Finally we demonstrate that to use this technique, an important procedure of multiple imputation needs to be followed on missing values, avoiding holes in the series that cause misinformation about the series. We provide an interesting step-by-step method that demonstrates its usefulness with real data.

Our goal is to identify patterns in the series data such as: identification of potential patterns due to trends in the series, identification of outliers for the prior knowledge about the structure of the time series and identification of potential patterns under the estimation and diagnosis of ARIMA models. As mentioned earlier, the assumption here is that the time series may be produced by an underlying model. The series components in a time-series mining task should provide enough insight into the prediction process. Hence, the goal is to find parameters for the model representation, whereas the task of prediction is explicitly to use the model. Therefore, a time series can be defined as a set of continuous time periods. A series is multivariate when several series simultaneously represent multiple dimensions within the same time span. This robust approach overcomes the issue that each record in a time series is unique. Firstly, from our analysis, we will focus on identifying general patterns due to trends in the series. The method at this stage includes plotting the series and its autocorrelation function to find out whether the series shows any upward or downward trend. An analysis of the series (see Figures 7 to 10) show that the persistence is more prevalent in some parts of the series than in others. Departures of the autocorrelations can be seen in those figures where horizontal lines show the series means. Tendency of highs to follow highs or lows to follow lows characterise series with persistence or positive autocorrelation or the other way around. A weak stationary linear behaviour seems to be present in both series because the line crosses the mean of the sample many times, thus increasing the chances for the variable to be stationary.

1) Imputing Series Data

Multiple imputation is a procedure that brings together the advantages of producing unbiased estimates and a suitable level of uncertainty in the model [22]. Imputing multiple series data for each missing observation should consider

correlation issues. Traditional approaches assume that the missing values are linear functions of other variables. However, this assumption misses a critical aspect in time series data, like the tendency of variables to move smoothly over time [15]. Our analysis ratified the fact that MCAR (missing completely at random) has been violated ($p<0.01$) which is common in real data mechanisms [21]. Then in particular for series data, a usual assumption is that the data points are MAR (missing at random). An overall data examination indicates 6.4% missing values. The missingness pattern in the BUN-Port dataset is given on both series *CoffOutBUN* and *CoffExpCol*. Little [21] suggests that the estimates of the imputation depend on the strength of the associations between variables in the dataset.

Honaker and King [15] show a simple but sophisticated approach to reach an accurate imputation for time-series data, and allow the incorporation of the usual assumption in multiple imputation: that the data are MAR. The assumptions considered in our imputation model are:

a) The complete data matrix denoted by $D(n,k)$ is multivariate normal, with n time points, k measurement variables, mean μ and covariance matrix Σ, that is $D_{(n,k)} \sim \mathcal{N}_k(\mu,\Sigma)$
b) The missingness matrix denoted by M indicates whether or not a value $m_{i,j}$ is missing for observation i and variable j. The usual assumption of MAR depends on the observed data D_{obs}: $p(M|D) = p(M|D^{obs})$
c) The likelihood of the observed data D^{obs} using the MAR assumption and the $D(n,k)$ parameters $\theta=(\mu,\Sigma)$ is $p(D^{obs}, M|\theta) = p(M|D^{obs})\, p(D^{obs}|\theta)$

Our imputation involves five multiple runs to produce results using R, Amelia II (Multiple Imputation of Incomplete Multivariate Data Analysis) and Zelig Package. The multiple imputation takes into consideration arguments such as polytime (polynomials included in the imputation to account for effects of time) and lags (indicating columns in the data that should have their previous values included in the imputation model). Table 2 shows the obtained parameters and statistics for each imputation for the whole series from 1999 to 2012. Results for *CoffOutBUN* showed that imputation 1 produced the best results since its mean (27692) and standard deviation (8615) were closest to the original mean (28485) with standard deviation (8275), while results for *CoffExpCol* showed that imputation 4 had the closest values (the original mean 45800 and the imputed mean 44902 with standard deviations of 11593, 11682 respectively).

	CoffOutBUN with df=163				CoffExpCol with df=164			
	μ	σ	*F*	*p*	μ	σ	*F*	*p*
1	27692	8615	23.5	1.04e-09	44193	11643	31.0	1.03e-07
2	27038	8765	26.6	1.01e-10	44830	11934	33.7	3.26e-08
3	27649	8427	20.2	1.46e-08	45414	11886	32.6	5.17e-08
4	27058	8622	26.6	1.01e-10	44902	11682	33.6	3.35e-08
5	27455	8421	23.5	1.03e-09	44946	11631	32.5	5.29e-0

Table 2. Statistics obtained by multiple imputation

2) Detection of Multivariate Outliers

Visualising frequency distributions allows for insight into the concentration of values. We classify our variables as continuous scale measurements, thus screening continuous variables for normality is an important step in multivariate analysis. In a univariate analysis of outliers, four cases were cleansed using Tukey's exploratory data analysis. In a multivariate analysis we also need to be aware of multivariate outliers which might be observed for certain combinations of the variables. Detecting a multivariate outlier is not a straightforward process. One method frequently used is Mahalanobis distance or the distance of a case from the centroid of the remaining cases in a multivariate space. Cases far from others appear to be multivariate outliers; a 1% significance level is used for identification of the outliers. For example in figure 4 a multi-outlier case 103 is found, therefore to improve the multivariate analysis, the case was deleted. A further analysis of outliers will show that in series analyses other kinds of outliers are detected; however, at that point they are left in the data to allow their impact on the series.

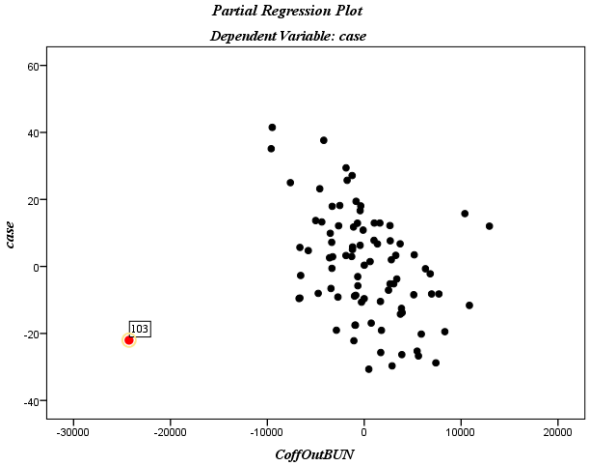

Figure 3. Identification of multivariate outliers

3) Identification of potential patterns due to series trend

Because the ARIMA cycle of identification, estimation, and diagnosis takes longer when seasonal processes are present, we seasonally adjusted the series before estimation. Therefore the assumption is that they are annual seasonally as expected in observations gathered monthly over the year; however, because we have shortened the whole series into five spans, seasonality is not easily readable. Moreover, real data even after seasonal adjustments exhibit trends, cycles, random-walking and other non-stationary behaviour. We will show further analyses in the following steps:

Step 1. Autoregression analysis: the basic step is to run an autoregression or linear autoregressive function on the observations considering *CoffExpColTon* and *CoffOutBUN* as dependent variables in function of time. Figures 4 and 5 present two plots with interesting patterns in the autocorrelation function (ACF). While figure 4 indicates a random walk presence, figure 5 strongly suggests the presence of a cyclical pattern.

Step2. Cross-correlation analysis: Darlington suggests examining potential patterns using the cross-correlation function for multivariate time series in contrast to controlling covariance such as in an ordinary least squares regression analysis (OLS). A negative lag indicates that the first series follows the second series. A positive lag for the spanned series 1999 to 2003 is obtained though at lag 0. Therefore, in this

case neither series can predict the second one with absolute certainty. For the spanned series 2001 to 2005, a weak cross-correlation at lag 11 is observed; however, a nonsignificant positive lag indicates that the first series specified: *CoffOutBUN*, leads the second series in this lag. That also means this series is a leading indicator of the second one and works best at predicting.

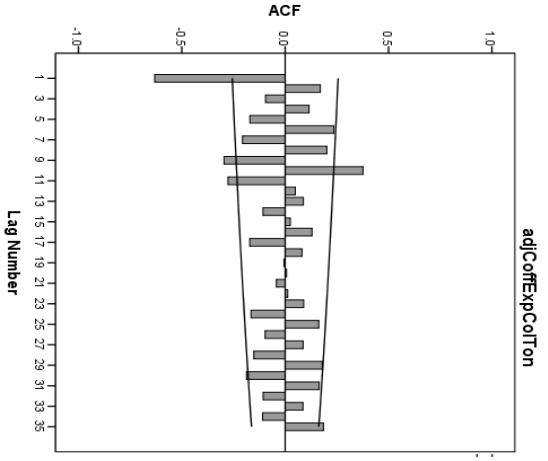

Figure 4. Interesting patterns in the ACF – Span 1999-2003

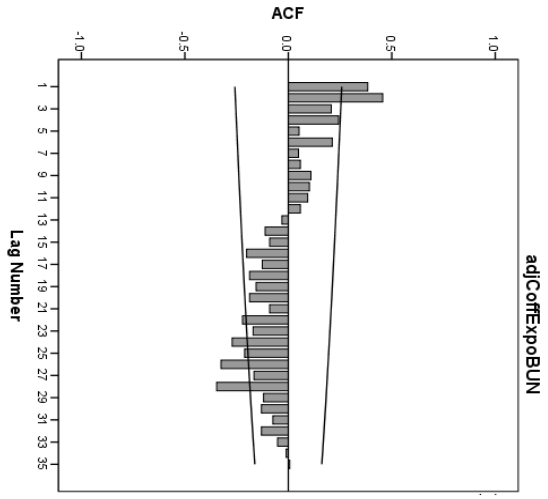

Figure 5. Interesting patterns in the ACF – Span 2001-2005

4) Identification of Outliers for the Prior Knowledge About the Structure of the Time Series

An important aspect of TSDM is to be able to make predictions under the presence of outliers. There is plenty of software able to perform automated detection of outliers such as IBM SPSS 19. What we pursue here as Milanovic and Stamenkvic [25] suggest, is a visual detection and valid characterisation of outliers that allows for the knowledge about the structure of the series even if they are seasonally adjusted. We attempt to classify those data points as hidden patterns that affect the changes in the series and may indicate some meaning [14]. Among others, Sánchez and Peña [30] identify possible types of outliers: additive and seasonal additive outliers (AO). The main difference between these two is that the former does not affect the subsequent observations but it affects a single observation. The latter occurs repeatedly at regular or seasonal intervals. Sánchez & Peña [30] also use other unified outlier patterns such as level shift (LS), transient level change (TC) and innovational (IO). If an outlier shifts many observations by a constant and has a permanent effect then may be considered LS. In contrast if it exponentially decays over the subsequent observations then it may be considered TC. Finally, an outlier is considered IO if it is additional to the noise term in the series. Table 3 displays the outlier analysis for each series and includes extra information such as if the case was imputed or not.

Process: Autoregression CofExpCol				**Process: Autoregression CoffOutBUN**			
case	span	type	imputed	Case	span	type	imputed
39:Mar02	99-03	AO	yes	04:Apr99	99-03	IO	no
49:Jan03	99-03	AO	yes	15:Mar00	99-03	AO	no
34:Oct03	01-05	AO	yes	44:Aug02	01-05	TC	no
				30:Jun03	01-05	TC	no
				51:Mar05	01-05	AO	no

Process: Forecast **Direction: CoffOutBUN → CoffExpCol**			
case	**span**	**type**	**imputed**
42-Jun04	01-05	AO	no
46:Oct03	01-05	AO	no

Table 3. Outliers' classification in time series analyses

The main conclusions we draw from Table 3 are: (a) case 15 (see top RHS) is a global minimum detected at the *CoffOutBUN* spanned series 1999 to 2003. It is clear that after this observation (i.e., AO) the downward trend turns upwards. The same kind of outlier is seen in the series *CoffExpCol* cases 39 and 49 (see top LHS). Imputation does not seem to be the cause for their presence because they happen even without imputed values. (b) An IO is observed in case 4 (see top RHS) indicating a local maximum. Moreover, we noted that it is the highest point for half of this period. Sanchez and Pena [26] demonstrate that an IO may have an effect on random walk processes, whereas an AO may affect the series correlation. Indeed we found a similar result when considering case 15 as a trigger effect for highs to follow highs.

5) Identification of potential patterns under the estimation and diagnosis of ARIMA models

In this part we present the model identification using appropriate ARIMA (p,d,q) and the seasonal equivalents (P,D,Q) in both series. This is accomplished under the use of ACF (autocorrelation functions) and PACF (partial autocorrelation functions). The summary of the results is given in Table 4. Four models can be distinguished: the first two are autoregressive individual series for the two spanned periods; the next two are the series analysis with explanatory or predictor variable; and those according to the best result obtained in each period. Notice the third model uses *CoffExpoColTon* as predictor, whereas the fourth model uses *CoffOutBUN* as predictor. Autoregressive models of order one or two AR(1) or AR(2), were found most suitable for the *CoffOutBUN* series. Whereas the integrated models I(1) to provide models for non-stationary through the differenced data, were most likely found in *CoffExpColTon* series. The general multiplicative ARIMA model for series with seasonality was found in both series as expected in observations gathered monthly over the year. This seasonal component was easily seen in the cycles inferred for the *CoffOutBUN* series.

After the ARIMA models' identification and fitting for each individual series, it was possible to cross-correlate the series to determine the best set of positive lags to predict the dependent variable through the predictor. In the diagnosis stage of developing of these models it is necessary to evaluate the stationary R-squared statistic (maximum fit value is 1) which provides an estimate of the proportion of the total variation in the series that is explained by the model [16]. A

Ljung-Box non-significant value of <0.05 means that there is sufficient evidence that the model is appropriate to describe the series. Selecting the single series provides a lower stationary R-squared in all cases in comparison with the selection of the series under predictor's influence. However, notice in Table 4 that the direction for this prediction is important. The spanned series 2001 to 2005 with the initial predictor *CoffOutBUN* for *CoffExpColTon* series, reported an interesting stationary R-squared of 0.787 and Ljung-Box non-significant value of 0.568; whereas the spanned series 1999-2003 reported unsatisfactory results in this direction. Thus we may conclude either the spanned series 1999 to 2003 are strongly affected by the four additive outliers reported above, creating a random walk process difficult to fit, or the prediction may be explained in both directions. The statistics for *CoffExpColTon* series as a predictor of *CoffOutBUN* in the period 1999 to 2003 are: stationary R-squared of 0.634 and Ljung-Box non-significant value of 0.636. An examination of residuals using the ACF and PACF plots accounts for the correct model specification and estimation of these models. No significant lag exceeds the boundaries of significance (i.e., a significant result suggests discrepancies between the model and the data). Next we discuss these results in the context of a domain application.

				Span 99-03			
Single Series	*n*	AR	I	MA	SAR	D	SMA
Model_1.CoffExpColTon	60	0	1	0	1	0	0
Model_2. CoffOutBUN	60	1	0	0	1	0	0
Series+Predictor	*n*	AR	I	MA	SAR	D	SMA
Model3. CoffExpColTon ***predicts*** *CoffOutBUN*	6	1	0	0	1	0	0
				Span 01-05			
Single Series	*n*	AR	I	MA	SAR	D	SMA
Model_1.CoffExpColTon	56	0	0	0	1	0	0
Model_2.CoffOutBUN	56	2	0	0	0	1	0
Series+Predictor	*n*	AR	I	MA	SAR	D	SMA
Model_4.CoffOutBUN ***predicts*** *CoffExpColTon*	56	0	0	0	0	1	0

Table 4. Summary of ARIMA Identification Models

IV. Discovering Patterns Driving an Underlying Forecasting Model

This section summarises the knowledge acquired in the previous steps. Our first learning was about the signatures of persistence in the series. From the initial sequence charts it was possible to identify segments of the series with high autocorrelation and weak seasonality. After this, an important connection between the initial patterns and the ACF and PACF functions gave enough detail to describe the trends of the series. A second learning was present in the embedded analysis of outliers. We noticed how the spanned series 1999 to 2003 was strongly affected by the reported four outliers (additive type - AO) creating a random walk process that was difficult to fit (*CoffExpColTon*). However, we also noticed that the prediction improves with the inclusion of the predictor that has a more stable trend (*CoffOutBUN*). A third learning is in connection with the domain. Thus we offer here an attempt to explain historical trends and the forecasts obtained:

In South America, particularly in countries like Colombia, a crisis is being experienced due to the lack of investment in infrastructure to carry out general improvements particularly to the seaports. Kent and Ashar [18] indicate that the ports of Colombia were privatized between 1990 and 1997. The period observed between 1999 and 2003 is thus a post-privatisation period. It is therefore feasible to assume that the national economy particularly in regards to the exports of coffee may be unstable. And so it is appropriate to treat *CoffExpColTon* as an independent variable. Forecasts for seven months ahead for 2004 are validated within the upper and lower ranges with 95% confidence. Analyses for the spanned series 2001 to 2005 encompass part of this trend. However, the changes in the economy may have marked improvement in seaport performance and operations (compared to what it was before); though the improvement in seaport management and technology has not been quite as impressive as in other countries around the world [2], [6], [11]). We see in figure 6 how the port performance indicator *CoffOutBUN* gains importance in the prediction of *CoffExpColTon*, thus indicating that the port operation was able to affect the economy in regards to the exports of coffee somehow by explaining 78.7% of its variation. Forecasts for five months ahead for 2006 are validated within the upper and lower ranges with 95% confidence.

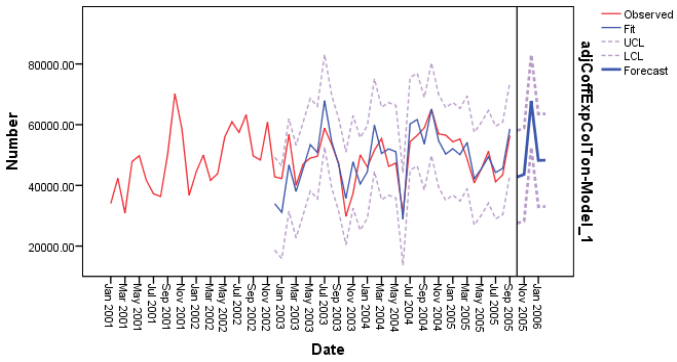

Figure 6. Fitting plot spanned series 2001-2005

V. The Interoperability of *i*-DMSS for Port Integration

Times Series	**CoffOut BUN**	
	Notation	Value
Length of the time series	L	60
Length of testing series	L_T	7
Yearly period	Y_p	1999-2003
Yearly testing period	Y_{Tp}	2004
Periodicity	P	Monthly
Type of the time series	*TYPE*	Micro-economy
Basic trend (non-seasonally adjusted)	*BT*	2.999
Basic trend (seasonally adjusted)	BT_{SAj}	1.475
Test of turning points	*TP*	14
Avg. coefficient of autocorrelation	*AC*	– 0.088
Percentage of missing values	mv.p	0.016 (1/60)
Percentage of missing values	mv.p	0.037
Subset ratio	set.r	0.3

Table 5. Meta-features Module III – Case 4: BUN Port South America

As stated by Tang and MacLennan [35, p. 38]: "It has been very difficult to integrate the result of data mining with user

applications to close the analysis loop. Most data mining products don't have APIs [Application Program Interfaces]. It is very painful to integrate data mining features with many business applications". Specific characteristics of a dataset can be extracted providing a common mining schema that can attain the robustness of parameters and accuracy in the modelling results. We followed the meta-features extraction, describing the series as presented in Tables 5 to 8. Most of the features are directly computed from the series data; however, a basic consideration for extracting the meta-feature: Test of turning points (TP) is that of examining the movements of the series. A number of peaks are evident crossing the series mean. Therefore the turning points are those at the highest or lowest positions in relation with the series mean. Figures 7 to 10 show evidence of the TP extraction.

Times Series	**ColExpColTon**	
	Notation	Value
Length of the time series	L	60
Length of testing series	L_T	7
Yearly period	Y_p	1999-2003
Yearly testing period	Y_{Tp}	2004
Periodicity	P	Monthly
Type of the time series	*TYPE*	Macro-economy
Basic trend (non-seasonally adjusted)	*BT*	3.946
Basic trend (seasonally adjusted)	BT_{SAj}	2.373
Test of turning points	*TP*	17
Avg. coefficient of autocorrelation	*AC*	–0.122
Percentage of missing values	mv.p	0.367 (22/60)
Percentage of missing values	mv.p	0.037
Subset ratio	set.r	0.3

Table 6. Meta-features Module III – Case 5: BUN Port, South America

This interoperability is possible through the predictive modelling mark-up language technology (PMML). The PMML technology allows a blend of several independent data mining solutions resulting in a PMML file containing multiple models. "This is the power of PMML: enabling true interoperability of models and solutions between applications. PMML also allows [shielding] end users from the complexity associated with statistical tools and models [13, p.5]". This subsection shows the standard procedure of using PMML in time series models. Guazzelli [12, 13] offers a standard in PMML 4.0 version that allows ARIMA models be supported in this technology.

Times Series	**CoffOut BUN**	
	Notation	Value
Length of the time series	L	60
Length of testing series	L_T	5
Yearly period	Y_p	2001-2005
Yearly testing period	Y_{Tp}	2006
Periodicity	P	Monthly
Type of the time series	*TYPE*	Micro-economy
Basic trend (non-seasonally adjusted)	*BT*	0.724
Basic trend (seasonally adjusted)	BT_{SAj}	0.543
Test of turning points	*TP*	11
Avg. coefficient of autocorrelation	*AC*	0.269
Percentage of missing values	mv.p	None
Percentage of missing values	mv.p	0.037
Subset ratio	set.r	0.3

Table 7. Meta-features Module III – Case 5: BUN Port, South America

Times Series	**ColExpColTon**	
	Notation	Value
Length of the time series	L	60
Length of testing series	L_T	5
Yearly period	Y_p	2001-2005
Yearly testing period	Y_{Tp}	2006
Periodicity	P	Monthly
Type of the time series	*TYPE*	Macro-economy
Basic trend (non-seasonally adjusted)	*BT*	4.356
Basic trend (seasonally adjusted)	BT_{SAj}	2.618
Test of turning points	*TP*	13
Avg. coefficient of autocorrelation	*AC*	0.036
Percentage of missing values	mv.p	0.416 (25/60)
Percentage of missing values	mv.p	0.037
Subset ratio	set.r	0.3

Table 8. Meta-features Module III – Case 5: BUN Port South, America

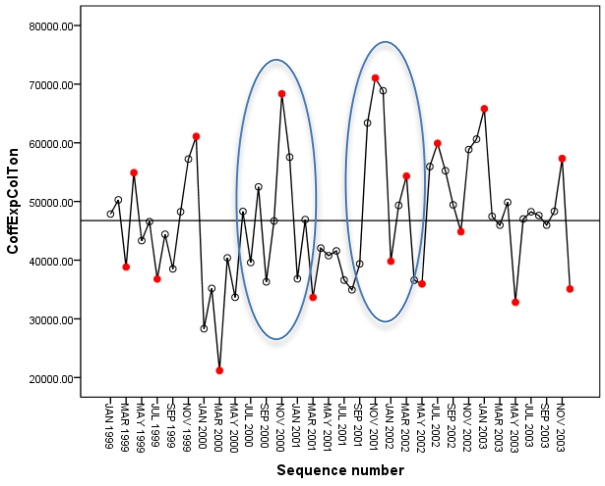

Figure 7. TP extraction – series *CoffExpColTon* 1999 to 2003

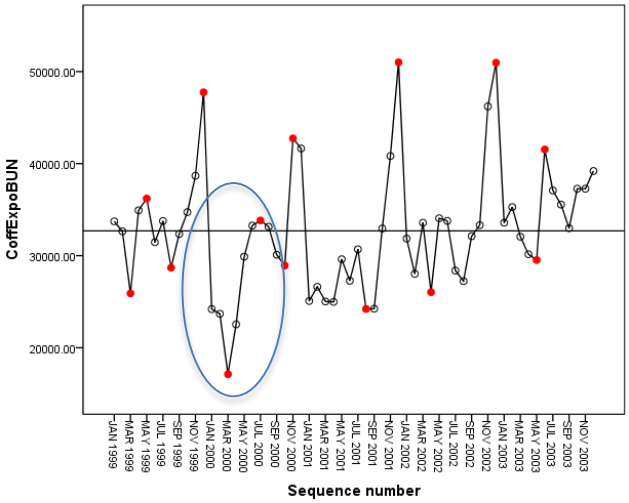

Figure 8. TP extraction – series *CoffOutBUN* 1999 to 2003

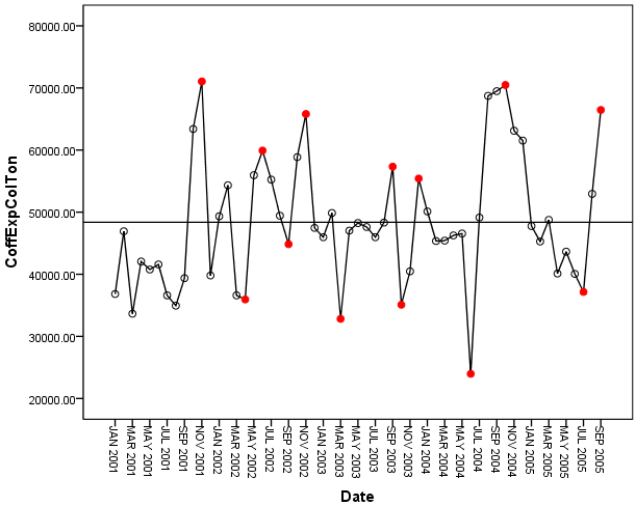

Figure 9. TP extraction – series *CoffExpColTon* 2001 to 2005

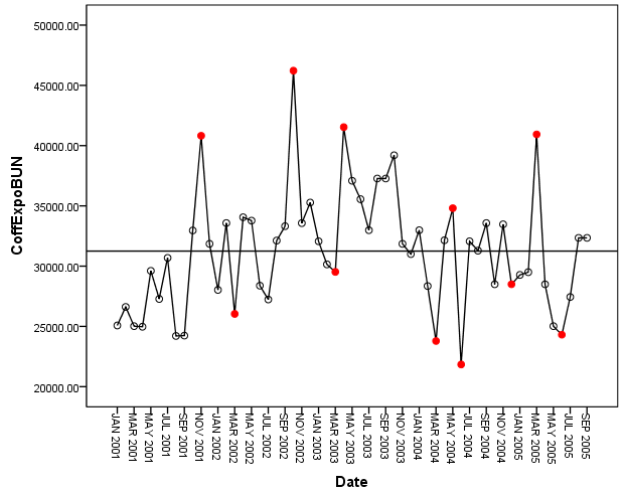

Figure 10. TP extraction – series *CoffOutBUN* 2001 to 2005

A time series model in PMML is represented by attributes such as:

- Year firstCaseIndex: a year time index from which all time measurements are made. In this case "2000".
- Month firstCaseIndex: a month time index from which all time measurements are made. In this case "0"
- CycleLength: the element containing the series sequence. In this case "12".
- SeasonLength: the number of points to be grouped using the previous CycleLength. In this case "12".
- Transformation function: necessary attribute transformations. In this case "none"
- PeriodLength: required to give the length of the season. In this case "57".
- LagTerm: accomplishes the identification process of ARIMA at different lags. In this case "11".
- PredictorEffect variableID: identifies the predictor variable to determine the predictive result in the dependent variable. In this case "adjCoffExpoBUN".

We also present below in Listing 1 an extraction of time series elements in PMML. The model is specified using ARIMA and the following elements. Listing 2 shows an extraction of the PMML code for the ARIMA specification. The PMML code also displays the parameter estimates and the significance tests based on the model selected. Sections IV and V in this paper provided a detailed explanation of ARIMA model selection. Two of these tests are predominantly used: stationary R-squared = 0.787 and the Ljung-Box non-significant value of 0.568. Listing 3 shows an extraction of the PMML code for the parameter estimates and the significance tests based on the model selected.

VI. Conclusions

Using a setup with one port from South America, Buenaventura port (Colombia), the element of port economic sustainability and operation efficiency in developing regions was included in the decision of port integration. This setup is generic and can be applied to any region in the world facing the decision of port integration. The interrelated activities to induce domain knowledge following Stabell's approach, i.e., data collection principles, descriptive modelling and normative modelling (norm of how decisions should be made), form the generic framework to support the creation of a DSS with a more accurate and machine interpretable representation called in this paper as the conceptual intelligent decision making support system (i-DMSS) for port integration.

The key objective of this work is to offer a model-based approach to TSDM based on the assumption posed by Esling and Agon [8] that the time series may be produced by an underlying model. We summarized the acquired knowledge in three basic learning processes: (a) learning about the signatures of persistence in the series (b) learning the meaning of observed outliers in the series and (c) learning the underlying model in connection with the domain. Interesting results indicate that for the period 2001 to 2005, the port performance indicator *CoffOutBUN* gains importance in the prediction of *CoffExpColTon*, thus indicating that the port operation was able to affect the economy in regards to the exports of coffee. The previous period was strongly affected by four outliers, creating a random walk process difficult to fit but feasible to produce due to unstable conditions evidenced in the economy. We offer this model as a solution for the question of how to develop forecasting models resulting from business processes that can be embodied in an intelligent decision support system. Though, limitations of this model may be envisaged for problems in which time series exhibit a more dynamic behaviour. As the system generating time series is not necessarily linear or stationary, the TDSM framework can build upon nonlinear and non-stationary time series. Those models cannot be produced using ARIMA assumptions.

Acknowledgement

This paper is an extended version of the work entitled "A Baseline Time Series Data Mining Model for Forecasts in Port Logistics and Economics", by A. Halabi, D. Richards, and A. Bilgin, published in the Proceedings of the 13th International Conference Intelligence Systems Design and Applications (ISDA 2013), pages 314-319, 2013.

References

[1] H. M Al-Deek,"Comparison of two approaches for modeling freight movement at seaports". *Jrnl of Computing in Civil Eng*., 15(4), 284-291, 2001.

[2] A. Baird. "Analysis of private seaport development: the port of Felixstowe". *Transport Policy*, 6(2), 109-122, 1999.

[3] I. Batyrshin, , & L. Sheremetov. "Perception Based Time Series Data Mining for Decision Making". in *Theoretical Advances and Applications of Fuzzy Logic and Soft Computing* (Vol. 42, pp. 209-219), O. Castillo, P. Melin, O. Ross, R. Sepúlveda Cruz, W. Pedrycz & J. Kacprzyk (eds.), Springer Berlin Heidelberg, 2007.

[4] K. Bichou, "An empirical study of the impacts of operating and market conditions on container-port efficiency and benchmarking. Research" in *Transportation Economics*, 42(1), 28-37, 2013. doi: http://dx.doi.org/10.1016/j.retrec.2012.11.009

[5] Celik, M., Lavasani, S. M., & Wang, J. "A risk-based modeling approach to enhance shipping accident investigation". *Safety Science*, 48(1), 18-27, 2010.

[6] K. Cullinane, W.-Y. Yap, & J.S.L. Lam. "The Port of Singapore and its Governance Structure". *Res.in Transp. Economics*, 17, 285-310, 2006.

[7] P. De Langen, M. Nijdam, M. van der Horst. "New indicators to measure port performance". *Journal of Maritime Research*., IV (1), 23-36, 2007.

[8] P. Esling, & C. Agon. "Time-Series Data Mining". *ACM Computing Surveys*, 45(1), Article No. 12. DOI: 1210.1145/2379776.2379788, 2012.

[9] P. Fourgeaud, "Measuring Port Performance. Measuring Port Performance". *The World Bank* (eds), (pp. 1-18), 2000.

[10] T. C. Fu. "A review on time series data mining". *Engineering Applications of Artificial Intelligence*, 24(1), 164-181, (2011).

[11] J.R.M. Gordon, P.-M. Lee & Jr. H.C. Lucas. "A resource-based view of competitive advantage at the Port of Singapore". *The Journal of Strategic Information Systems*, 14(1), 69-86, 2005.

[12] A. Guazzelli, W. Lin & T. "Jena PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics". (2nd ed.). Charleston, SC, 2013.

[13] A. Guazzelli. "What is PMML? Explore the power of predictive analytics and open standards", 2010. Retrieved from http://public.dhe.ibm.com/software/dw/industry/ind-PMML1/ind-PMML1-pdf

[14] J. Han, & M. Kamber. "Data Mining: Concepts and Techniques" (2nd ed.). San Francisco: Morgan Kaufmann Publishers, 2006.

[15] J. Honaker & G. King. "What to Do about Missing Values in Time-Series Cross-Section Data". *American Journal of Political Science*, 54(2), pp. 561-581, 2010.

[16] IBM SPSS. "Forecasting 20", 2011 Retrieve from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Forecasting.pdf

[17] M. Kadous & C. Sammut. "Classification of Multivariate Time Series and Structured Data Using Constructive Induction". *Machine Learning*, 58(2-3), 179-216, 2005.

[18] P. Kent and A. Ashar. "Indicators for port concession contracts and regulation: the Colombian case", 2010. Retrieve from http://www.asafashar.com/IAME_2010_Article_Performance_Indicators_for_Regulators_Final_Final.pdf

[19] D. Kroenke, D. Bunker, & D. "Wilson. Experiencing MIS" (3rd ed.): Pearson Education, Inc, 2012

[20] W. H. K. Lam, P. L. P. Ng, W. Seabrooke, & E. C. M. Hui. "Forecasts and reliability analysis of port cargo throughput in Hong Kong". *Journal of Urban Planning and Development-Asce*, 130(3), 133-144, 2004

[21] T. Little. "Longitudinal structural equation modelling". The Guilford Press, NY, 2013.

[22] L. Littvay. "Corruption and democratic performance". University of Nebraska-Lincoln, 2006

[23] I. Lopez-Yanez, L. Sheremetov, & C. Yanez-Marquez. "A novel associative model for time series data mining". *Pattern Recognition Letters*, 41, 23-33. doi: 10.1016/j.patrec.2013.11.008, 2014.

[24] M. Magala. "Targeting market opportunity for port growth: a CART-based decision support system". *Maritime Policy & Management*, 34(2), 131-150, 2007.

[25] M. Milanovic, & M. Stamenkovic-Radak. "Data Mining in Time Series". *Ekonomski Horizinti*, 13(1), 5-25, 2011.

[26] D. H. Ni, J. D. Leonard, & Trb. "Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data". *Information Systems and Technology* (pp. 57-67), 2005

[27] J. C. Pomerol, & F. Adam. "Understanding Human Decision Making - A Fundamental Step Towards Effective Intelligent Decision Support". in *Intelligent Decision Making: An Ai-Based Approach* (Vol. 97, pp. 3-40), G. PhillipsWren, N. Ichalkaranje & L. C. Jain (eds.), Berlin: Springer-Verlag Berlin, 2008

[28] R. J. Povinelli. "Time Series Data Mining: Identifying Temporal Patterns for Characterization and Prediction of Time Series Events". (Doctoral dissertation), Faculty of the Graduate School of Marquette University, Wisconsin, United States, 1999

[29] C. Ratanamahatana, J. Lin, D. Gunopulos, E. Keogh, M. Vlachos, & G. Das. "Mining Time Series Data". in *DM & KD.Handbook* (pp. 1049-1077), O. Maimon & L. Rokach (eds.), Springer US.

[30] M. J. Sanchez, & D. Pena. "The identification of multiple outliers in ARIMA models". *Communications in Statistics-Theory and Methods*, 32(6), 1265-1287, 2003.

[31] C. Stabell. "A Decision-Oriented Approach to Building DSS". in *Building Decision Support Systems* (pp. 221-260), J. L. Bennett (eds.).Reading, MA: Addison-Wesley, 1983.

[32] W. Talley. "An Economic Theory of the Port". *Research in Transportation Economics*, 16, 43-65, 2006a.

[33] W. K. Talley. "Chapter 22 Port Performance: An Economics Perspective". *Research in Transportation Economics*, 17(0), 499-516, 2006b.

[34] W. K. Talley. "Performance indicators and port performance evaluation". *Logistics and Transportation Review*, 30(4), 339-352, 1994.

[35] Z. Tang & J. MacLennan. "Data Mining with SQL Server". *John Wiley & Sons*, 2005.

[36] S. P. Tiwari. "Development of ports in Saurashtra and Kutch region: an economic analysis". PhD thesis, Saurashtra University, 2011. Retrieve from

http://shodhganga.inflibnet.ac.in/handle/10603/3984?mode=full

[37] M. Tull & F. Affleck. "The Performance of Western Australian Ports", 2007. Retrieved from http://www.patrec.org/publication_docs/31_Regulation%20of%20WA%20ports%20FINAL%20June07.pdf

[38] X. Wang, K. Smith, & R. Hyndman. "Dimension Reduction for Clustering Time Series Using Global Characteristics". in *ICCS* (Vol. 3516, pp. 792-795), V. Sunderam, G. Albada, P. A. Sloot & J. Dongarra (eds.), Springer Berlin Heidelberg, 2005.

[39] B. Zondag, P. Bucci, P. Gutzkow, & G. de Jong. "Port competition modeling including maritime, port, and hinterland characteristics". *Maritime Policy & Management*, 37(3), 179-194, 2010.

Author Biographies

Ana Ximena Halabi Echeverry is a Professor in the School of Economics and Management Sciences at Universidad de La Sabana, Chia, Colombia. She has been awarded with the international Macquarie Research Excellence Scholarship (iMQRES) to complete the Ph.D. on Computing Sciences in Macquarie University Australia and was recently Visiting Scholar at the United Nations University Institute on Comparative Regional Integration Studies (UNU-CRIS). Her research interests lie broadly in Data Mining and Knowledge Discovery, Modelling Expertise, Supply Chain and Informational Integration. She has published in academic peer-reviewed journals such as IEEE Publications and Springer Special Editions and has served as a reviewer of papers for various academic journals. Follow her work at www.researchgate.net

Deborah Richards is a Professor in the Department of Computing at Macquarie University. She joined academia in 1999, following 20 years in the IT industry during which she completed a BBus (Comp and MIS), MAppSc (InfoStudies) and PhD in artificial intelligence on the reuse of knowledge. While she continues to work on systems that assist decision making and knowledge management, for the past decade, her focus has been on agent technologies including agent-based modelling, the use of agent based virtual worlds and empathic virtual agents to improve engagement, adherence and support for students and patients to deliver improved learning and/or health outcomes.

Ayse Bilgin Ayse Bilgin is a Senior Lecturer in the Department of Statistics at Macquarie University in Sydney, Australia. Her research interests include statistics education and applied statistics especially in health sciences. She teaches undergraduate and postgraduate students in various topics such as Operations Research, Data Mining and recently she developed the Capstone unit – Statistical Consulting - for the students majoring in Statistics. Her work has been presented in various Statistical education conferences, health related conferences and published as refereed conference proceedings and/or journal articles.

Jairo R. Montoya-Torres is Full Professor in the School of Economics and Management Sciences at Universidad de La Sabana, Chia, Colombia, and Director of the Operations & Supply Chain Management Research Group at the same university. He was recently Visiting Scholar at Leeds University Business School, University of Leeds, UK, under a Marie Curie International Incoming Fellowship of the Seventh European Framework Programme. He obtained the Postdoctoral Degree for Research Direction (Dr.-Hab.) from Institut National des Sciences Appliquees (INSA) de Lyon and Universite Claude Bernard, Lyon France in 2011, and a PhD from Ecole National Superiere des Mines de Saint-Etienne and Universite Jean Monnet, Saint-Etienne, France, in 2005. His research interests lie broadly in supply chain management under collaborative and sustainable environments, simulation and optimisation of logistics and production systems, production scheduling and vehicle routing. He has been a member of several academic societies, and has served as Guest Editor or within the Editorial Board of various international academic journals. His personal web page is http://jrmontoya.wordpress.com.