Structural time series: computational efficiency in estimating economic parameters in industry

J.A. Fiorucci†, M.G. Andrade², D. Nascimento³, L. Ferreira⁴, A. Leite⁴, L. Cavalcante⁵, D. Frota⁶, O. Egbon⁶, M.S. Bressiani⁷, F. Louzada²

¹ Department of Statistics, University of Brasilia, Brasilia, Brazil
² Institute of Mathematical Sciences and Computing, University of São Paulo, São Carlos, Brazil
³ Universidad de Atacama, Copiapó, Chile
⁴ National Institute for Research in Computer Science and Automation, France
⁵ School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, Brazil
⁶ Federal Institute of Education, Science and Technology of Ceará, Sobral, Brazil
⁷ Center for Mathematics, Computing and Cognition, Federal University of ABC, São Paulo, Brazil

(Communicated to MIIR on 25 August 2021)

Study Group: VI Brazilian Study Group with Industry - Mathematics in Industry Reports (MIIR), Cambridge Open Engage

Communicated by: Francisco Louzada Neto and José Alberto Cuminato

Industrial Partner: Kinea-Itaú https://www.kinea.com.br

Presenter: Lucas Tumkus de Lima and Alexandre Cristovao Maiorano

Team Members: A. Leite, National Institute for Research in Computer Science and Automation, France; D. Frota, Federal Institute of Education, Science and Technology of Ceará, Sobral, Brazil; D. Nascimento, University of Atacama, Atacama, Chile; F. Louzada, Institute of Mathematical Sciences and Computing, University of São Paulo, São Carlos, Brazil; J.A. Fiorucci, Department of Statistics, University of Brasilia, Brasilia, Brazil; L. Cavalcante, School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, Brazil; L. Ferreira, Institute of Mathematical Sciences and Computing, University of São Paulo, São Carlos, Brazil; M.G. Andrade, Institute of Mathematical Sciences and Computing, University of São Paulo, São Carlos, Brazil; M.S. Bressiani, Center for Mathematics, Computing and Cognition, Federal University of ABC, São Paulo, Brazil; O. Egbon, Institute of Mathematical Sciences and Computing, University of São Paulo, São Carlos, Brazil;

Industrial Sector: Data Analysis; Finance;

Tools: Statistical Models, Dynamic Regression, ARIMA, Stepwise, Bayesian Network

Key Words: Structural Time Series, Dynamic Regression Models, ARIMA, Stepwise, Bayesian Network

MSC2020 Codes: 91B84, 37M10, 62M10, 62P20, 82C03

† Corresponding Author: jafiorucci@unb.br
Summary

Under the Industrial 4.0 and Big data revolution, observational data-driven models have often been used to support decision-making and causal relations. Nonetheless, probabilistic thinking is guaranteed in the small data world if correct suppositions are attended to. A growing field is related to automatized Time Series analysis, through complicated due to the dependence of observed and hidden dimensions often presented in these data types. In this report the problem is motivated by a Brazilian financial company interested in unraveling relation structure explanation of the Japanese’ CPI ex-fresh Food & Energy across 157 economical exogenous variables, with very limiting data. The problem becomes more complex when considering that each variable can enter the model with lags of 0 to 8 periods, as well as an additional restriction of admitting only a positive relationship. This report discusses three possible treatments involving models for structured time series, the most relevant approach found in this study is a Dynamic Regression Model combined with a Stepwise algorithm, which allows the most relevant variables, as well as their respective lags, to be found and inserted in the model with low computational cost.

1 Introduction

Computational demand has been one of the greatest challenges in research fields that work with numerical simulation and big data mining to solve today’s problems, even more whenever the acquired data shows some structural dependence (e.g. time or spatial).

In the finance and the economic team, it is often common to map economic variables and understand the factors that can affect/impact the market (in anticipation). Additionally, nowadays, the statistical modeling combined with computational power is done through exhaustion tests, series association using multiple linear regressions, considering variable selection across all the possibilities in the set (variables and their lags). In other words, it is considered in a set of variables its optimal subset, seeking for the model with the best explanation or predictability power.

A Brazilian finance company proposed a problem related to structural time series modeling with a series of challenges and constraints to be solved within a week during the workshop.

This report describes the approach that was proposed to solve the problem and some directions for other possible solutions. The next two subsections present the problem in details and a summary of what was proposed. Section 2 details the methodology used in this study, while the results are presented in section 3. Finally, the conclusion of this study is presented in section 4.

1.1 The problem

The proposed challenge consisted of finding out an optimized computational model, enabling to explain the economic variables "Japanese CPI ex-fresh food & energy" through-
out some structural model identified over a high-dimensional economic data set with 157 variables, which can be considered with lags ranging from 0 to 8, thus totaling 1413 possible variables. The estimation procedure considers a relatively small data set sample formed by 116 quarterly observations.

They aimed to extract as much information as possible, however with some restriction such as lag order up to eight periods by series and just positive relationships. Besides, the lagged dependent variables shall not be in the model as covariable (feature/predictor). These restrictions were included because of the a priori treatment that the company applied to the database and so that the results could be more easily interpreted by company analysts.

It is worth noting that the main objective of the presented problem is to explain the oscillations in the time series of interest as a function of the covariates, so that the methodology require variables selection to explain an economic cycle. In this case, forecasting is usually not a priority.

Some variables are chosen to map the country’s annual aggregated power variation, for instance, the mortgage gap tax, the total annual wage variation, and some others, thus enabling the analysts to follow the consumption evolution changes.

The flux connected to this process is pretty limited due to the computational demand. Consequently, the analysts are pushed to create new variables, choose their lags, and select them, limiting the forecasting.

Another difficulty in this kind of model is the lack of economic variables interrelation since the estimation construction is not vectorial. Besides, it is challenging to handle causality, co-integration, and multicollinearity in this process’s linear model.

In summary, the presented problem includes the following challenges:

- Identification of structural relationships in time series;
- High-dimensional and few observations data set;
- Selection of a predefined number of variables (ten) to be included in the model;
- Only one lag per selected variable is allowed;
- Presence of strong correlations in some variables;
- Simple interpretation modeling;
- Previously treated variables;
- Only positive relations are allowed;
- Computationally fast and scalable solution to other similar problems;

1.2 Team Approaches

This report targeted the problem of high-dimensional structural time-series data (more explainable variables than observations) in section 2.1, which propose the use of a Step-wise algorithm combined with a Dynamic Regression Model. This solution meets all the requirements of the problem presented in section 1.1 and presents very interesting results in terms of computational efficiency and goodness of fit, these results are presented in section 3.2.

In addition to this solution, other approach have also been proposed which partially address the Problem 1.1. Given the high-dimensional characteristic of the problem, a
traditional approach to regression problems that require variable selection is the use of LASSO-regulated estimation, which uses an L1 regularisation penalty to achieve sparsity in a regression. However, this is a traditional approach for linear regression problems that do not involve auto-correlation in the variables, so its application to time series, although used in some works, does not have strong scientific support. LASSO forces some variables’ contribution to zero and seems to be a good strategy when combined with Bayesian Network representation given an understanding of the variables correlation structure. In this manner, we adopted the Bayesian Network with LASSO-regularization to shrink the parametric space through its variance-covariance matrix and then adopt a Bayesian Structural Time Series approach, which is present in section 2.2.

2 Theoretical Foundation

2.1 Dynamic Regression Model

When we model economic time series, the effect of the past observations and external variables’ effect may be relevant. In this sense, Dynamic Regression Models [1] are appropriate since we can use linear regression to incorporate the information from the covariates.

Suppose $X_1, X_2, \ldots, X_k$ is a set of observed random occurrences, such as explanatory variables. An observed response variable (Y) can be written in terms of regression coefficients $\beta_i, i = 1, 2, \ldots, k$ as

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,t} + \eta_t, \quad (2.1)$$

thus, the $\{Y_t\}$ process is decomposed into a linear regression structure. However, in our case $\{Y_t\}$ is a time series and therefore the traditional assumption $\{\eta_t\}$ being an independent process is not valid, fortunately this problem can be addressed by assuming that $\{\eta_t\}$ is an Autoregressive Moving Averages (ARMA(p,q)) process, which is given by

$$\eta_t = a_1 \eta_{t-1} + a_2 \eta_{t-2} + \cdots + a_p \eta_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + \cdots + b_q \varepsilon_{t-q}, \quad (2.2)$$

where $\{\varepsilon_t\}$ are equally distributed and independent with $\varepsilon_t \sim N(0, \sigma^2)$ and $a_1, \ldots, a_p$, $b_1, \ldots, b_q$ are parameters to be estimated along with $\beta_0, \ldots, \beta_k$.

The Dynamic Linear Regression Model formed by the equations (2.1) and (2.2) is well know in literature and used called ARMAX or LM-ARMA, the traditional reference to this model is Box et al. [2].

The equation (2.2) can be written as a function of the delay operator $(B)$ and then compressed as

$$\Phi(B) \eta_t = \Theta(B) \varepsilon_t,$$

where $\Phi(B) = 1 - a_1 B \cdots - a_p B^p$ is called Autoregressive Polynomial of order $p$ and $\Theta(B) = 1 + b_1 B \cdots - b_q B^q$ is called Moving Average Polynomial of order $q$.

An important hypothesis of this modeling is that all variables $(Y, X_1, \ldots, X_k)$ are stationary processes, in this way it could be shown that the error component $\{\eta_t\}$ is also stationary. For this property to be contemplated by the model, a restriction to the
autoregressive parameters \((a_1, ..., a_p)\) must to be included, the ARMA process will be stationary only if the roots of the \(\Phi(B)p\) are outside the unit circle \([2]\).

It can be shown that this model can be written as a linear state space model with Gaussian innovations \([1]\), from which it proves that the maximum likelihood estimator of the set of all parameters corresponds to the least squares estimator, which can only be obtained numerically by computational optimization.

An alternative to use this model for non stationary variables is to prior differentiate these variables to remove unit-roots \([1]\). Thus, the first step in applying the model should be the verification of stationary of each variable and the respective application of differentiation when necessary.

Then, in a (invariant) structural time series problem the second step is to select the variables \((X’s)\) with the respective lags that best explain the variable of interest \(Y\). This step can be performed by testing all possible combinations, however, the computational cost tends to be high, which can make this practice unfeasible in high-dimensional data sets. In this case, the literature \([1]\) proposes that this selection should be performed using a stepwise algorithm, which considers the contribution of each variable added to the model according to a parsimony criterion, traditionally the Akaike Information Criterion (AIC) \([3]\) or its corrected version for small samples (AICc) \([4]\).

A procedure for selecting variables and lags according to all the requirements of Problem 1.1 is proposed below. Such methodology presents several advantages that we can use, easy interpretability, easy implementation, simplified potential, and interval forecasting capabilities as long as the covariates’ future is known.

**LM-ARMA STEPWISE**

To select the ten best features in the data set, we developed a stepwise methodology. In this methodology, we assume that all the dataset variables are stationary since we will calculate the correlation between time series and also fit stationary models. The steps we follow in this stepwise feature selection are described below.

**Step 1:** Calculate the correlation of the response variable and each lag of the other variables in the data set. In Problem 1.1 was required to use lags between 0 and 8.

**Step 2:** Select the lag of each variable with the greatest correlation with the response variable.

**Step 3:** Create a data set with the lag of each variable that has the greatest correlation with the response variable and sort the variables in descending order of correlation.

**Step 4:** Fit a regression with stochastic modeling in the response variable using, as a covariate, the feature that had the greatest correlation. This feature will be the first variable to be selected in the stepwise method.

**Step 5:** Determine the number of features to select. In Problem 1.1 was required to select ten features.

**Step 6:** Add to the model the other variables in the dataset (one by one until 10) in the order that they were ranked in Step 3. If by adding a new variable, it improves the AICc (decreases its value) then it is kept in the model, otherwise (if the value of the AICc increases) then it is discarded. In addition, any additional restrictions related to variables can be included in this step, as for example the restriction of Problem 1.1 of
only having a positive relation in the model, so if when including a new variable the sign of the coefficient of some variable becomes negative, then it must be discarded.

The presented algorithm has low computational cost, since the presented stepwise procedure has an order of complexity equal to the number of covariates \((k)\). This algorithm was proposed in contrast to the conventional stepwise (forward) algorithm, which has a quadratic order of complexity \((k^2)\).

### 2.2 Bayesian Network

Probabilistic graphical models can be represented as a Bayesian Network (bn). It can be obtained by considering the joint probability distribution as a product of independent conditional functions over a set of variables. The gain of a bn representation is the visualization associated with the powerful tool, representing the uncertain knowledge of a given domain and is depicted as the nodes of a network, as a dag.

The joint probability distribution over all variables is computed as the product of all these conditional probabilities dictated by the arcs. This distribution entails enough information to attribute a probability to any event expressed with the network variables. Moreover, efficient algorithms for computing any such probability without having to generate the underlying joint probability may be unfeasible in many cases. bn have enormously progressed over the last few decades leading to applications spanning all fields \([5]\).

For instance, \(n\) variables \(X = (X_1,...,X_n)\) are ruled by a domain that is of great interest since it contains all the information. Furthermore, they can be used to ask many probabilistic questions. Moreover, they can be decomposed by using the chain rule such as \(P(X_1,...,X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\cdots P(X_n|X_1,...,X_{n-1})\). Likewise, the conditional independence between the variables can be exploited, enabling us to reduce the joint probability function into a compact expression. For instance, two random variables \(X\) and \(Y\) are conditionally independent given a random variable \(Z\) if 

\[
P(X|Y,Z) = P(X|Z) \quad \forall x,y,z \in X,Y,Z. 
\]

In other words, whenever \(Z = z\) is observed, to explain \(X = x\), all the information is held in \(Z\). Therefore the information \(Y = y\) does not influence the probability of \(x\). The random variables \(X,Y,Z\) can even be disjoint random vectors. In this case, their definitions can be written as 

\[
P(X,Y|Z) = P(X|Z)P(Y|Z). 
\]

In this manner, conditional independence is found whenever one can find for each \(X_i\) in a set of parent \(Pa(X_i) \subseteq \{X_1,...,X_{i-1}\}\) such that given \(Pa(X_i)\). Then \(X_i\) is conditional independent of all the variables in \(\{X_1,...,X_{i-1}\}\) if \(Pa(X_i)\). As a result, 

\[
P(X_1,...,X_n) = P(X_1|Pa(X_1))\cdots P(X_n|Pa(X_n)) 
\]

A graphical representation can be used to represent the dependence between the variables. In this case, a graph \(G\) comprises as a pair \((V,E)\), where \(V\) is the set of nodes and \(E\) is the set of edges between the nodes in \(V\). This factorization of the joint probability with a dag representation is known as bn. The next topic will discuss how to deal with a small sample size with a large number of parameters to estimate the bn.
Structural time series: computational efficiency in estimating economic parameters in industry

2.2.1 LASSO

Estimating bn features encounter severe limitations as the number of parameters to estimate grows quickly with the network’s size. In practice, to estimate a bn with many parameters, the number of observations needs to exceed the number of samples in the data set. lassolasso [6] has been used to with this problem.

lasso regression is an example of regularized regression. Let us suppose a regularized regression model on a data set with \( n \) observations and \( m \) features.

\[
Y = \beta_0 + \sum_{j=1}^{m} X_{ij} \beta_j
\]

LASSO regression is an \( L_1 \) penalized model where we add the \( L_1 \) norm of the weights to our least-squares cost function:

\[
J(\beta) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{j=1}^{m} |\beta_j|
\]

lasso employs a regularizing penalty by limiting the total sum of absolute parameter values. It treats positive and negative edge-weights equally, leading to many edge estimates shrinking to zero, and as a result, being drop out of the model. As such, the LASSO returns as sparse (or, in substantive terms, conservative) network model: only a relatively small number of edges are used to explain the covariate structure in the data. Because of this sparsity, the estimated models become more interpretable. The lasso utilizes a tuning parameter to control the degree to which regularization is applied. This tuning parameter can be selected by minimizing the ebic [7].

2.2.2 Bayesian Structural Time Series Model

The Bayesian Structural Time Series Model (bsts) model is a Bayesian alternative to the ARIMA model. It comprises a powerful tool to handle uncertainty in data more firmly and cleverly. Unlike the ARIMA, bsts model representation does not include lags, difference, and moving averages of the observed variables, and its structure fully determines it. bsts models have been applied in various fields such as in management technology [8], financial market [9], ecology [10], health policies [11], among others.

Let \( Y_t \) represents a time series occurrence over a time period \( t \in N \). A general form of a bsts model with regression term can be formulated as

\[
Y_t = Z_t^T \mu_t + x_t^T \beta + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2), \\
\mu_{t+1} = T_t^T \mu_t + R_t^T v_t, \quad v_t \sim N(0, H_v^2), \tag{2.3}
\]

where, \( \mu_t = (\tau_t, s_t)^T \) is a vector of the unobserved trend \( \tau_t \) and seasonality \( s_t \) whose evolution is defined by bsts. The model vectors \( Z, T, \) and \( R \) typically contain a mix of zeros and ones. In this study, \( Z_t = (1,1)^T, T_t = R_t = (1,0)^T, \) and \( v_t = (v_t,0). \) The error terms \( \epsilon_t \) and \( v_t \) are assumed to be serially and mutually independent, with hyper-parameters \( \sigma_\epsilon^2 \) and \( H_v^2. \) By our specification, \( \sigma_t^2 = \sigma_\epsilon^2 \), and \( H_v^2 = \sigma_v^2 \). The vector \( x_t = (x_{1t}, ..., x_{pt})^T \) is a \( 1 \times p \) vector of linear regressors with associated coefficient vector \( \beta = (\beta_1, ..., \beta_p)^T. \)
Table 1. Variables and lags from the linear model provided by the company to explain the “Japanese CPI ex-fresh food & energy” (Y)

| Variables                      | lag | Estimative |
|-------------------------------|-----|------------|
| intercept                     | -   | 0.060      |
| (X1) Dummy VAT hike           | 0   | 2.113      |
| (X2) Dif OutputGap            | 3   | 0.235      |
| (X3) Import Price             | 8   | 0.005      |
| (X4) Brent YoY                | 3   | 0.001      |
| (X5) Dif. Will Durables       | 2   | 0.001      |

bsts is fully specified by setting a prior distribution on \( \beta \). For large \( p \), the cost of estimation may become unbearable. Consequently, a shrinkage prior may be appropriate to perform a variable selection on the covariates \( x_t \) that explain the time series observation. This analysis assigned a Spike and Slab priors on \( \beta \). The R package bsts utilizes mcmc to estimate the parameters.

3 Results

In this section we present the application of the proposed methodology to the data set provided by the company. At first, in section 3.1, for comparison purposes, the use of a multiple linear regression (LM) model will be presented, which is known to be not an approach recommended in the literature. In section 3.2, we present the application of the methodology proposed and discussed in section 2.1.

For computational implementation was used the R language [12] and the forecast package [13, 14]. The data set and code are available at https://github.com/jafiorucci/6WSMPI-Problem-4.

3.1 Why should not use a simple linear regression model to this problem?

Since the main objective of Problem 1.1 is to decompose the series of interest (Y) according to the explanatory variables (X1, X2, ...) in order to have an easy interpretation, a first approach that may seem convenient is to use a simple linear regression model, this would be equivalent to assume the model presented in equation (2.1) has \( \{\eta_t\} \) as an independent and identically distributed Gaussian process, \( \eta_t \sim N(0, \sigma^2) \).

In fact, this approach was considered by the company, which after testing several possibilities to combine five explanatory variables that met the pre-defined requirements, arrived at a linear model where the variables are presented in Table 1. The fitted values in contrast with observed values are presented in Figure 1.

Figure 2 show the residual of this model suggesting a remained pattern, also supported by Ljung-Box test (p-value =2.22e-16), rejecting independence across the \( \{\eta_t\} \) process.

Since the independence assumption was violated, according to Chapter 9 of Hyndman and Athanasopoulos [1] the model will present several problems, in particular the coefficients estimated by this model are not the best and possibly unreliable. Furthermore, statistical tests of significance are not valid in this situation.
3.2 Dynamic regression model with a stepwise approach

The algorithm presented in section 2.1 was implemented in R and applied to the dataset. In fact, it turned out to be quite computationally efficient and feasible to scale to more complex problems, as the database from Problem 1.1 was processed in approximately three seconds on a laptop equipped with an Intel i7 9750H processor and a Windows 10 operating system.

The stationarity of the response variable "Japanese CPI ex-fresh food & energy" was verified using the KPSS test which found no evidence to reject the stationarity hypothesis (p-value=0.1). The same test was performed for all explanatory variables at the 5% significance level, for which it was applied to differences until stationarity was obtained, none of which required more than one difference. Seasonal unit root was also verified, but it was not found for any variable. The autocorrelation graphs (ACF and PACF) shown in Figure 3 suggest a non-seasonal ARMA(p, q) model with low values of p and
Table 2. Selected variables with respective lags, Maximum Likelihood Estimators (MLE), Standard Errors (SE), Statistics (Z) and p-values

| Variables                       | lag | MLE  | SE   | Z    | p-value |
|---------------------------------|-----|------|------|------|---------|
| ar1                             | -   | 0.83 | 0.07 | 12.35| 0.00    |
| ma1                             | -   | 0.33 | 0.11 | 2.92 | 0.00    |
| ma2                             | -   | 0.32 | 0.11 | 2.91 | 0.00    |
| intercept                       | -   | 0.27 | 0.17 | 1.58 | 0.11    |
| (X1) Dummy VAT hike             | 0   | 1.45 | 0.10 | 13.89| 0.00    |
| (X2) OutputGap                  | 2   | 0.06 | 0.03 | 2.32 | 0.02    |
| (X3) JP I.P. - CEMENT & CEMENT PROD. VOLA | 7 | 1.44 | 0.91 | 1.59 | 0.11    |
| (X4) JP I.P. - COSMETICS VOLA   | 1   | 1.17 | 0.30 | 3.97 | 0.00    |
| (X5) JP I.P. - MOTORCYCLES VOLA | 5   | 0.26 | 0.18 | 1.45 | 0.15    |
| (X6) JP I.P. - FOODS AND TOBACCO VOLA | 1 | 2.64 | 0.77 | 3.45 | 0.00    |
| (X7) JP I.P. - LEATHER GOODS VOLA | 5 | 2.29 | 0.69 | 3.30 | 0.00    |
| (X8) JP I.P. - CERAMICS, CLAY & STONE P. VOLA | 5 | 0.80 | 0.39 | 2.06 | 0.04    |
| (X9) JP I.P. - ELECTRONIC COMPUTERS VOLA | 8 | 0.46 | 0.22 | 2.06 | 0.04    |
| (X10) LME-SHG Zinc 99.995% 3 Months US/MT | 8 | 0.20 | 0.07 | 2.84 | 0.01    |

$q$, thus, the algorithm presented in section 2.1 was tested with $p = 1$, $p = 2$, $q = 0$ and $q = 1$. The AICc criterion was used to select the ARMA orders and variables, which found the LM-ARMA($p = 2$, $q = 1$) with the variables listed in Table 2 as the most appropriate configuration. Table 2 also presents the stepwise selected variables and lags, their parameter estimates and significance tests.

Figure 3. Graphics of Auto-correlation Functions for "Japanese CPI ex-fresh food & energy" variables

Figure 4 shows the fit of the LM-ARMA model (red curve) and the observed data (black curve). Observing Figure 4, we can see the excellent approximation between the fitted and observed curves that the model provides a satisfactory fit to represent the observed data.

Figure 5 presents the model residual diagnoses, and as can be seen, the autocorrelation function (ACF) of the model residuals is very similar to white noise ACF. The selected model summarizes the information available on the data set aims to explain the time
Figure 4. Observed time series and fitted time series model by the selected model

Figure 5. Residual Diagnostics of the selected model

series analyzed. In Figure 5 (a) the observed residuals show stationary, in Figure 5 (b) can be observed no significant correlation between residuals and in Figure 5 (c) can also be observed the excellent agreement of this series histogram and a histogram of the variable with normal distribution. The QQPlot (Figure 6) reinforces our claim. As can be seen, the model residuals approach very well the characteristics necessary for a hypothetical normal distribution.

Table 3 shows that the null hypothesis of stationarity, independence, and normality of the residuals are not rejected. KPSS test resulted in a p-value greater than 0.1, indicating no rejection of the null hypothesis of stationary residuals. Ljung-box test also resulted in a high p-value, indicating no rejection of the null hypothesis of independence. Shapiro-
Wilk provided the same indication of no rejection but now of the null hypothesis of normal distribution residuals.

Accordingly, with the obtained results, it is possible to observe the model’s good adequacy and residuals, enabling us now to estimate non-biased parameters and test their significance. The correlation between the selected explanatory variables are presented in Figure 7, which, as expected, shows that the stepwise algorithm used did not select variables with a high level of correlation (as it would harm parsimony), thus it is reasonable to assume that the variables used do not present collinearity.

Figure 6. QQplot of the residuals of the selected model

Table 3. Statistical test of the residuals

| Test          | $H_0$       | Statistic | P-value |
|---------------|-------------|-----------|---------|
| KPSS Stationarity | 0.07        | > 0.1     |
| Ljung-Box Independency | 10.89 | 0.95    |
| Shapiro-Wilk Normality  | 0.99      | 0.75      |

Figure 7. Correlation in the selected explanatory variables
4 Conclusion

The modeling initially presented by the company has some theoretical controversies when treating time series data (and therefore auto-correlated) as independent variables. This model was tested throughout this report, which, as expected, presented residuals that do not satisfy the model required hypothesis.

This report proposes the use of a Dynamic Regression Model, which is a methodology similar to the linear model used by the company, but which deals with auto-correlation in the error component, thus being suitable for time series data. Furthermore, to select the explanatory variables, as well as their respective lags, an algorithm was proposed that first builds a rank of the explanatory variables and lags that are more correlated with the response variable and then runs a stepwise-type algorithm to include these variables up to a predetermined limit. This algorithm allowed the entire variable selection process to be carried out with low computational cost. The use of the stepwise algorithm also has the advantage of minimizing the collinearity problem.

The residual analysis of the proposed model showed satisfactory results and fully aligned with the model’s hypotheses. In this way, the relationship between the variables that are revealed by the model can be considered consistent. Furthermore, statistical tests could also be performed to verify the significance of the variables, which at a significance level of 5% revealed eight significant variables of a total of ten variables that were selected and included in the model.

Preliminary results from the Bayesian Network are presented and discussed in the Appendix.
References

[1] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[2] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[3] H. Akaike, “A new look at the statistical model identification,” *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[4] C. M. Hurvich and C.-L. Tsai, “A corrected akaike information criterion for vector autoregressive model selection,” *Journal of time series analysis*, vol. 14, no. 3, pp. 271–279, 1993.

[5] C. Bielza and P. Larranaga, “Bayesian networks in neuroscience: a survey,” *Frontiers in computational neuroscience*, vol. 8, p. 131, 2014.

[6] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[7] J. Chen and Z. Chen, “Extended bayesian information criteria for model selection with large model spaces,” *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[8] S. Jun, “Bayesian structural time series and regression modeling for sustainable technology management,” *Sustainability*, vol. 11, no. 18, p. 4945, 2019.

[9] O. Poyser, “Exploring the determinants of bitcoin’s price: an application of bayesian structural time series,” *arXiv preprint arXiv:1706.01437*, 2017.

[10] N. Droste, C. Becker, I. Ring, and R. Santos, “Decentralization effects in ecological fiscal transfers: A bayesian structural time series analysis for portugal,” *Environmental and Resource Economics*, vol. 71, no. 4, pp. 1027–1051, 2018.

[11] N. Chami, Y. Li, S. Weir, J. G. Wright, and J. Kantarevic, “Effect of strict and soft policy interventions on laboratory diagnostic testing in ontario, canada: a bayesian structural time series analysis,” *Health Policy*, vol. 125, no. 2, pp. 254–260, 2021.

[12] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[13] R. J. Hyndman and Y. Khandakar, “Automatic time series forecasting: the forecast package for R,” *Journal of Statistical Software*, vol. 26, no. 3, pp. 1–22, 2008. [Online]. Available: https://www.jstatsoft.org/article/view/v027i03

[14] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O’Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, and F. Yasmeen, *forecast: Forecasting functions for time series and linear models*, 2021, r package version 8.15. [Online]. Available: https://pkg.robjects.com/forecast/

[15] S. Epskamp, D. Borsboom, and E. I. Fried, “Estimating psychological networks and their accuracy: A tutorial paper,” *Behavior Research Methods*, vol. 50, no. 1, pp. 195–212, 2018.
Appendix: Bayesian Network Preliminary Results

In the following subsection, we will present an alternative to variable selection, using a graphical representation approach through Bayesian reasoning. Although this approach is promising, more investigations are needed in order to fully address Problem 1.1.

Bayesian Nets for Correlation Estimation

All the variables from the provided dataset, 157 variables, were lagged up to order 4, resulting in 778 variables. Then, we first adopted the BN with the LASSO regularization to estimate the potential variables explaining the Japanese CPI ex-fresh food & energy. This methodology was adopted due to the possibility of estimating the matrix of covariance (with sparsity representing the non-influence across some variables).

After removing the variables correlated greater than 0.95, the BN was estimated using the EBICglasso regularization [15] with tuning=0.5. Figure 8 shows the graphical representation of the correlation structure obtained from the BN method.

Figure 9 shows the centrality standardized measurement of each node. The descriptive metric enables one to understand more about the importance and connectivity of each node of the network, translating the high-dimension spaced relation into a single number.

Thus, given the restriction given by the company, which allowed to model the TS structure dependence up to 10 covariables, we selected the TOP 10 explain variables and adjusted a Bayesian Structural, as described in the following.

Figure 8. Estimated Bayes Net (BN) using EBICglasso.
Bayesian Structural TS model

The BSTS model was fitted on the BN top 10 selected variables as regressors to the Japanese CPI ex-fresh food & energy. Figure 10 shows the posterior estimates form the BSTS model, that is, the mean for each time point of the calculated components ("Trend", "Seasonality", "Regression"). Whereas the top panel is the smooth-trend of the time generating process. The middle panel represents the triennial estimated seasonality present in the process, while the bottom panel is the regression contributions (TOP 10 BN variables) to the generating process.

Figure 11 shows the bar plot of the average contribution of the non zero regression effects on the best model obtained from MCMC posterior samples. The height of the bars is the influence of the non-zero covariate effect on the time series. The most influential variable is "X.25".

The benefits of modeling Japanese CPI ex-fresh food & energy with the BSTS model are that it relaxes stationarity’s assumption in the time series process. Thus, it is not necessary to perform Lag difference.

To perform a model goodness-of-fitting visualization (BSTS versus LM), Figure 12 shows the residuals from these models and Table 4 some metrics associated as well.
Figure 10. BSTS estimated decomposition of the Japanese CPI ex-fresh food & energy.

Figure 12. Residual (error) from BSTS and LM, compared to the observed Japanese CPI ex-fresh food & energy.
Figure 11. Non-zero Variables from the BSTS model.

Table 4. Metrics of goodness-of-fitting from LM, BSTS and LM-ARMA.

| Model     | MAE  | MSE  |
|-----------|------|------|
| LM        | 0.373| 0.211|
| BSTS      | 0.281| 0.193|
| LM-ARMA   | 0.146| 0.034|