Statistical Model Selection for Seasonal Big Time Series Data

S

Brian Guangshi Wu
Southwestern University, USA

Dorin Drignei
Oakland University, USA

INTRODUCTION

Seasonal time series abound in areas such as environmental sciences and economics. For example, seasonal temperatures (e.g., Chen et al., 2016; Murthy & Kumar, 2021), seasonal precipitation (e.g., Sayemuzzaman & Jha, 2014; Martin et al., 2020) or seasonal wind speed (e.g., Shih, 2021) are common in environmental sciences, while seasonal business cycles (e.g., Gregory & Smith, 1996) or seasonal labor data (e.g. Liebensteiner, 2014) can be found in economics. Analyzing such data sets provides useful insight into seasonal patterns that have an impact on human activities and economic development. Due to recent capabilities to collect large amounts of data, however, classical statistical analysis methods have limitations, the big time series data sets posing new computational and modeling challenges.

In time series analysis, order identification refers to the selection of a time series model characterized by non-negative integer orders, which is followed by parameter estimation, diagnostic checking and forecasting. Despite being a critically important early step in time series analysis, order identification is perhaps the least developed among these steps. In autoregressive AR(p) processes the partial autocorrelation function is zero after lag p, thus identifying the AR order p. The autocorrelation function of moving average MA(q) processes is zero after lag q, providing a convenient method to identify the MA order q. The least-squares type method ESACF using the extended sample autocorrelation function has been proposed in Tsay and Tiao (1984) for the order identification of mixed autoregressive moving average ARMA(p,q) models. This method uses a sequence of linear regression models to identify the orders (p,q). A related method called SCAN has been proposed by the same authors in Tsay and Tiao (1985), using a canonical correlation approach. The applicability of these methods is facilitated by tables from which the orders p, q are identified. These methods can also be used for integrated ARMA (i.e. ARIMA) models. However, these methods are not directly applicable to other time series models, such as seasonal autoregressive integrated moving average (SARIMA) models, or certain types of nonlinear models. Cross-validation for time series model selection is a potential alternative, but it may be nontrivial to apply due to the inherent serial correlation (Bergmeir et al., 2018).

The most commonly used method for time series model selection is based on evaluating an information criterion for a few plausible time series models and choosing the model that minimizes such a criterion (e.g. Brockwell & Davis, 2016; Shumway & Stoffer, 2017). When a small set of plausible models is not available, one performs an exhaustive computation and minimization of the information criterion over a large enough grid of orders (Brockwell & Davis, 2016). However, choosing the best model using this exhaustive method is computationally challenging for big time series data, which could be a univariate large-sample time series (e.g. appliances energy consumption time series of length 19,735 in Candanedo

DOI: 10.4018/978-1-7998-9220-5.ch182

et al., 2017), a collection of large-sample time series (e.g. 438 stocks over 1,495 days in Lunde et al., 2016), or it can occur in a less common format (e.g. a temporal sequence of 606 facial expressions in Xu et al., 2020; Wang et al., 2020).

A method for the order identification of big time series data was developed in Wu and Drignei (2021), which was applied to ARMA models and ARMA-GARCH models, where the latter part stands for generalized autoregressive conditionally heteroscedastic. The goal of the current paper is to help time series practitioners better understand how to apply the method outlined in Wu and Drignei (2021) to SARIMA models of seasonal big time series. Here we focus on univariate large-sample seasonal time series, and the method can be summarized as follows. We compute the information criterion for a random sample of orders for the SARIMA model and use kriging-based methods to emulate (i.e. approximate) the information criterion for any new SARIMA orders. Then we use an efficient global optimization (EGO) algorithm to minimize the emulated information criterion, thus identifying efficiently the orders of SARIMA models for seasonal big time series. Both simulations and real temperature time series will be used to illustrate the method.

BACKGROUND

This paper focuses mainly on the challenges of statistical model selection for seasonal big time series, and potential solutions to address such challenges. However, such data sets have also generated much interest recently in other areas, such as machine learning. It is useful to review some recent works in such areas, while acknowledging that they discuss aspects of seasonal big time series data other than statistical model selection. Bachechi et al. (2022) incorporated information visualization methods into a system that can detect trends, seasonality, or anomalies in traffic flow over space and time. Castan-Lascorz et al. (2022) proposed a new prediction algorithm for time series, both univariate and multivariate, that uses clustering, classification and forecasting techniques. First, windows of time series values with similar patterns are grouped using clustering. Subsequently, for each pattern a prediction model is constructed only from the time windows associated with the pattern. The method can handle a large variety of time series behaviors, including seasonality. Guo et al. (2021) developed a hybrid model that combines the traditional time series model with an artificial neural network to predict the monthly mean atmospheric temperature profile. Carlini et al. (2021) proposed a method using graphs over time to study seasonal and trending patterns in world-wide maritime applications. Yi et al. (2021) used a penalized regression with inferred seasonality module (PRISM) and online internet search data to forecast unemployment initial claims. Poussin et al. (2021) used satellite-derived annual and seasonal time series of normalized difference water index to study implications of changing climatic conditions. Kumari and Toshniwal (2021) investigated a new hybrid deep learning model for global horizontal irradiance forecasting, which takes into account the spatio-temporal features of the data set. Wang et al. (2021) used a classification model that integrates the terrain, time series characteristics, priority, and seasonality with satellite images. Hewamalage et al. (2021) performed an empirical study to investigate the seasonal forecasting properties of recurrent neural networks. Da Silva et al. (2020) developed time series classification methods to monitor an agricultural area, based on active learning (AL) that selects limited seasonal time series information to generate the training set. Chen et al. (2019) proposed a Periodicity-based Parallel Time Series Prediction (PPTSP) algorithm for large-scale time series data showing periodicity characteristics. Jurman (2019) used a database including a large number of seasonal time series to predict the results of sport matches and competitions. Yeh and Yeh (2019) used queries on mortality related terms from Google