MIDAS regression using inflation and unemployment to predict GDP

Doga Bilgin Dina Jankovic Alexander Lam

March 2018

1 Introduction

Time series are essential in statistics, economics, quantitative finance, engineering, and many other areas. Time series analysis is used in many applications, such as economic forecasting, budgetary analysis, stock market analysis, process and quality control, inventory studies, etc. A time series could be defined as a sequence taken at successive equally spaced time points. The main goals of their analysis are forecasting, anomaly detection, clustering, classification, and query by content.

Researchers normally use classical linear models for time series analysis, with equal frequencies for all variables. Sadly, many important macroeconomic indicators are not sampled at the same frequency, and this may be an issue since standard forecasting models require equally spaced time intervals. However, Mixed Data Sampling (MIDAS) is an approach that can resolve this issue. It is a relatively unexplored econometric regression or filtering method

developed by Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov in 2004, where the independent variable(s) appear more frequently than the dependent variable. Their work focused primarily on volatility predictions, but has also proven to be useful for macroeconomic modeling.

The MIDAS models have been used in recent literature, such as in Clements and Galvão (2008) and Marcellino and Schumacher (2010), to improve the accuracy of predictions of quarterly GDP with monthly indicators for the United States and Germany, respectively. More recently, the specific usage of financial data paired with the MIDAS model to forecast GDP growth in the United States has been explored in Andreou et al. (2013). In short, these articles have concluded that the use of mixed frequency data significantly improves forecast accuracy.

Limitations to data availability is often a problem; for example, many research centres in the world cannot provide monthly information about the Gross Domestic Product (GDP), but they can for the Cost of Imports (CIM). They can provide the quarterly information about GDP, so there are missing values for GDP. A natural question is coming up – how do we examine the relationship between the high and low frequency variables? One possibility would be to calculate the arithmetic mean of the observations that occur between the lower frequency samples. This approach would imply equal slopes on each of the individual observations, which is an assumption that may be violated in many cases. For instance, more recent data are usually more informative. In this case, recent data should be assigned a higher weight

than earlier data. A simple linear regression using each daily value of the predictor variable as an individual regressor would require estimating a large number of parameters, thus leading to high estimation uncertainty.

Another motivating example is the following regression model for the risk-return trade-off:

$$R_{t+1} = \mu + \gamma \hat{\sigma}_t^2 + \varepsilon_{t+1} \tag{1}$$

where R_{t+1} is the excess return on the market in month $t+1$, and $\hat{\sigma}_t^2$ is the forecasted variance of returns for the same month $t+1$, based on information know at time t.

Normally, extra values for the high-frequency variable(s) will be available after the most recent sample value of the low-frequency dependent variable has been observed. In this case, these extra observations can be used as well, so we have the potential for what is usually termed "now-casting" in the forecasting literature. The hope is that incorporating this extra high-frequency information will improve the forecasting performance of the model.

The MIDAS approach allows regressors with different sampling frequencies and are therefore not autoregressive (AR) models, since the notion of autoregression implicitly assumes that data are sampled at the same frequency in the past. Instead, MIDAS regressions share some features with distributed lag models but also have unique novel features. The distributed lag model is a regression of the following form:

$$Y_t = \beta_0 + B(L)X_t + \varepsilon_t \tag{2}$$

where $B(L)$ is a finite or infinite lag polynomial operator, usually parameterized by a small set of hyperparameters.

In this paper, our goal is to study and implement the MIDAS regression model in order to predict the GDP using data from the Bureau of Economic Statistics from year 2009. The paper is organized as follows. In Section 2, we provide a detailed explanation of the methodology. We provide a description of the MIDAS regression model, together with its underlying probability distributions and parameters involved in the equation. In Section 3, we give a description of the data set used in this study. We explain why it is necessary to seasonally adjust the data and the possible ways to do it. We also explain the importance of stationarizing a time series, as well as the possible presence of unit roots and their detection. In Section 4, we provide the R code used in our analysis, together with the results.

2 Methodology

A simple forecasting model of the economy involves Ordinary Least-Squares (OLS) regression of GDP on its time lags, as well as on another related economic variable such as consumer prices, unemployment, or stock prices, and the related variable's lag:

$$Y_t = \beta_0 + \sum_{i=1}^{p} (\beta_i Y_{t-i} + \gamma_i X_{t-i}) + \epsilon_t \tag{3}$$

where Y_t is GDP, X_t is the related economic variable, and ϵ_t is an error term of zero mean and constant variance at time t. We can use data up to time t to predict the value of Y at time $t+1$.

However, time series data is often available at different frequencies. For example, the dependent variable, Y, is only published every quarter, whereas X is often higher frequency. CPI and unemployment data are available three times as often as GDP at monthly frequency, and stock prices are available intra-day. The simplest and most common method of dealing with different frequencies is to aggregate or average data so that all variables are at the lowest frequency. In this case, the quarterly averages of the high-frequency variable would be used as the regressor X in Equation (3). This method is equivalent to a restricted Least Squares regression where the coefficients on high-frequency lags of X are equal within the same quarter, which is the low-frequency period. However, such a model may not be ideal, as it essentially ignores the additional information provided by higher-frequency availability of certain data.

An alternative way to incorporate the high-frequency data is to simply regress Y on its own low-frequency lags as well as all of the high-frequency lags of X over the same horizon:

$$Y_{t^{LF}} = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t^{LF}-i} + \sum_{k=1}^{n*m} \gamma_k X_{t^{HF}-k} + \epsilon_{t^{LF}} \tag{4}$$

where m is the number of high-frequency periods contained in each low-

frequency period, and n is the number of low-frequency periods for which we want lags of X. Low and high frequency time periods are denoted as t^{LF} and t^{HF}, respectively. For the remainder of the methodology discussion, we will drop the constant and lag-dependent-variable terms from the regression equation, as they do not change across model specifications.

While Equation (4) allows for unique coefficient estimates for every single high and low frequency observation, there exist potential complications due to the proliferation of regressors. For example, if there are 60 daily stock price observations for every quarter, then a GDP forecasting model with only two quarter lags would already have 123 parameters to be estimated. This results in low degrees of freedom especially for time series without a long history, and it makes statistical inference imprecise. Furthermore, if close-together high-frequency lags are strongly correlated, the model will also suffer from multicollinearity problems. In order to reduce number of parameters, within-quarter effects could be restricted to be the same across all quarters:

$$Y_{t^{LF}} = ... + \sum_{j=1}^{n} \tau_j (\sum_{k=1}^{m} \gamma_k X_{t^{HF}-k})_{t^{LF}-j} + \epsilon_{t^{LF}} \tag{5}$$

However, in the stock price example, even this specification would require 65 parameters to be estimated for only two quarters of lags.

2.1 MIDAS regressions

The Mixed-data sampling (MIDAS) regression model is proposed by Ghysels, Santa-Clara, and Valkanov (2004) as a way to preserve information from high-frequency data while addressing parameter proliferation and multicollinearity problems. The MIDAS regression is based on a distributed lag model; instead of estimating the coefficient on each lag of the high-frequency regressor, the model instead assigns weights to the lags according to a polynomial function. The basic MIDAS model is as follows:

$$Y_{t^{LF}} = ... + \gamma \sum_{k=1}^{n*m} f(k;\boldsymbol{\theta}) X_{t^{HF}-k} + \epsilon_{t^{LF}} \tag{6}$$

where $f(k;\boldsymbol{\theta})$ is a polynomial function of the lag number k, and $\boldsymbol{\theta}$ is a small set of hyperparameters which govern the shape of the function. Because the lags are weighted according to the function, the coefficient estimates on the weighted lags are then restricted to an equal value γ. Thus, the MIDAS model can be estimated with far fewer parameters than in Equations (4) or (5), i.e. only γ and $\boldsymbol{\theta}$.

The function $f(k;\boldsymbol{\theta})$ can in practice be any type of polynomial function. Two common functional forms are the beta formulation and the exponential Almon function. Graphs of these functions from Armesto, Engemann, and Owyang (2010), based on selected values for two hyperparameters θ_1 and θ_2, are shown in Figure 1.

Based on these common functions, weights tend to decline as the high-

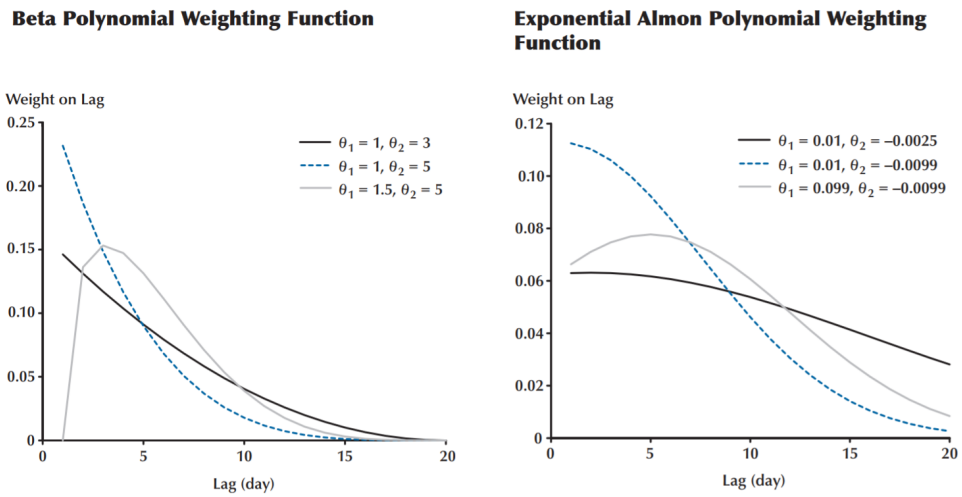

Figure 1: Lag weighting functions for MIDAS model

frequency lag increases, possibly with a hump-shape in more recent periods. The intuition behind the overall downward slope is that the effect of regressor data will likely diminish as it moves further into the past; however, recent movements in the regressor may take also take some time to pass through to movements in the dependent variable, leading to the hump-shape.

A useful feature of MIDAS regressions is that they can be used to forecast the dependent variable in the current low-frequency period using the availability of the high-frequency regressor data; this practice is known as now-casting. We add a term to the basic MIDAS model which incorporates the high-frequency lags of the regressor between the present and the beginning of the current low-frequency period. Thus, for the d^{th} high-frequency period in the current low-frequency period, the regression is as follows:

$$Y_{t^{LF}|d} = ... + \gamma_1 \sum_{j=m-d+1}^{m} f(k; \boldsymbol{\theta_1}) X_{t^{HF}+1-j} + \gamma_2 \sum_{k=d+1}^{n*m} f(k; \boldsymbol{\theta_2}) X_{t^{HF}-k} + \epsilon_{t^{LF}} \quad (7)$$

The additional term in Equation (7) means that effects of data from the current low-frequency period are estimated separately from previous periods using γ_1 and $\boldsymbol{\theta_1}$. Equations (7) and (6) are equivalent in the case that $\gamma_1 = \gamma_2$ and $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$.

3 Data

We apply MIDAS modelling techniques to economic data by showing how a country's Gross Domestic Product (GDP), a quarterly time series variable, can be forecast using monthly variables such as consumer prices and unemployment. We perform this exercise for data on the United States from 1980Q1 to 2017Q4.

Our data on real GDP in chained 2009 dollars comes from the Bureau of Economic Statistics at a quarterly frequency. In contrast, all of our explanatory variables are available at a monthly frequency. From the Bureau of Labor Statistics, we take the urban Consumer Price Index (CPI), excluding food and energy, as our measure of core consumer prices, as well as the civilian unemployment rate.

3.1 Seasonal adjustment

Often, time series data will exhibit seasonal patterns which repeat every year but are irrelevant for policy analysis purposes. These patterns may reflect the seasonal weather cycle or annual holidays. For example, GDP tends to be lower in the fourth quarter of the year as production stops for the Christmas holidays, while unemployment also tends to be lower due to seasonal hires. If we try to model the two time series without seasonal adjustment, the model may wrongly suggest a positive relationship between GDP and unemployment. In fact, this relationship is being driven by seasonal phenomena such as Christmas and as such the effect is not economically relevant. This is an example of omitted variable bias and may cause spurious correlation.

It may therefore be useful to seasonally adjust data such that all the effects of seasonal patterns are removed. Two common methods of seasonal adjustment are X-12-ARIMA and X-13-ARIMA, developed by the US Census Bureau. Broadly speaking, these methods use moving averages to decompose time series data into seasonal, trend, non-seasonal cycle, and idiosyncratic components. The seasonal component is then extracted out of the data. For the exercise in this paper, most of our data comes from US statistical agencies and is already seasonally adjusted at the source. The sole exception is for consumer price data, which we seasonally adjust ourselves using X-13-ARIMA.

3.2 Stationarity and first-differences

When doing time series analysis, we also want to ensure that all variables are stationary, meaning that they have a constant mean and variance over time. A non-stationary time series may possess a deterministic or natural time trend, or it may have a unit root. Roots are parameters of the variable's underlying autoregressive (AR) process that determine how the variable moves based on past values. Given a variable's equilibrium "steady state" value, a root equal to one, known as a unit root, means that movements of away from the equilibrium will become permanent so that there is no time-invariant mean. In contrast, a root between zero and one means that the variable will converge back to the equilibrium value over time, resulting in a constant mean. If a time series has a root greater than one, also known as an explosive root, it will also be non-stationary; however, this is an uncommon phenomenon in economic data.

Non-stationary time series can be spuriously correlated. Seemingly significant relationships can in actuality be explained either by a natural time trend, or simply because the variables exhibit similar AR processes and therefore move together. For example, a regression of infant mortality rates on the population of endangered whales would produce positive coefficient estimates. However, this relationship exists only because both variables are non-stationary and happen to both be declining; killing more whales would certainly not prevent infant deaths, nor would killing babies save the whales.

There exist common statistical tests which the presence of a unit root.

These tests either treat the unit root as the null hypothesis, such as the Dickey-Fuller test, or as the alternative hypothesis, such as the KPSS test. Non-stationary variables, whether or not they possess unit roots, can usually be made stationary by taking the first difference of the data, or the first difference of the natural log, depending on whether there is an exponential time trend. For the exercise in this paper, we make our data stationary by taking the first difference of the natural log of GDP and core CPI, and the first difference of the unemployment rate stock prices.

4 Nowcasting real GDP

4.1 Method

Many time-sensitive policy decisions rely heavily on infrequent data. Often, policy-makers obtain these essential data with a lag. Moreover, this data can be subject to revisions. Finally, some data is released at different times and at different frequencies (for example, data on inflation and unemployment are released on a monthly frequency while GDP is released on a quarterly frequency). Generating now-casts of key economic variables based on more readily available data serves as a useful exercise to perform policy decisions.

To illustrate an application of the MIDAS model, we develop a now-casting model to examine current levels of real GDP. Real GDP is a key input to many economic decisions, such as monetary policy, government budgeting

and business planning. However, real GDP is released on a quarterly basis in most countries. On the other hand, other indicators of economic activity, such as industrial production, unemployment, inflation, and stock prices are released at a monthly (or higher) frequency.

To evaluate the predictive power of our model, we compare the forecasts made by the MIDAS model to a regular OLS using a simple weighted average of the high-frequency variables. The model is estimated iteratively using an out-of-sample, rolling window forecast. This is done as follows. first, the model is estimated starting in 1980Q1 to the quarter preceding the nowcast quarter. The nowcast is then generated using the observations of the high-frequency variable in that quarter. The nowcast is then compared to the actual value using the root mean squared error (RMSE) [1].

4.2 Unemployment

The graph below compares the nowcasts generated using the unemployment rate for real GDP generated using MIDAS and using OLS. The actual real GDP reading in red, the MIDAS forecast in green and the OLS forecast in blue. In this example, the MIDAS model and OLS model seem visually similar. Comparing the RMSEs, the MIDAS model provides a marginal gain of about 5% over the OLS model.

[1] RMSE is a measure of model accuracy which equals the sum of the squared differences between the model's predicted values for the dependent variable and the actual realized values. $RMSE = \sqrt{\frac{\sum_{i=1}^{N} \hat{y}_i - y_i}{N}}$

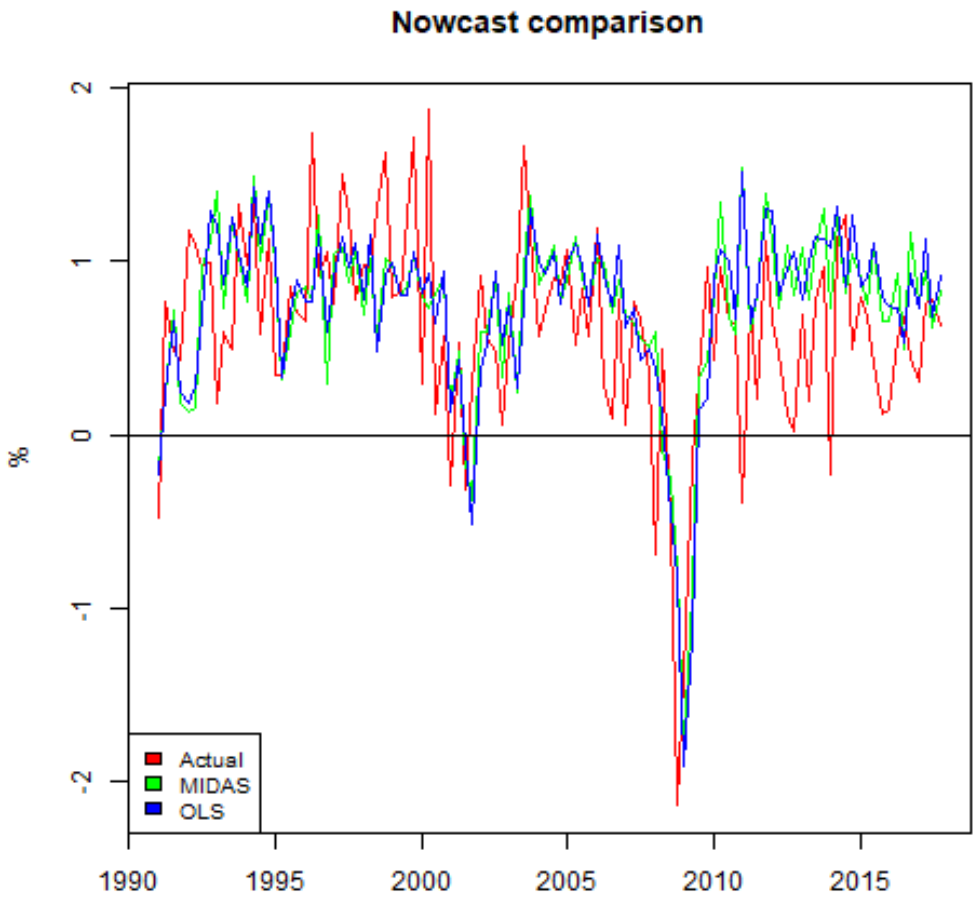

Figure 2: Nowcasts using unemployment

4.3 Inflation

Using the MIDAS model with seasonally-adjusted inflation as the explanatory variable, on the other hand, provides a much better nowcast for GDP than OLS. The chart below shows that while the OLS and MIDAS

model are broadly similar in most periods, the MIDAS model tends to better predict declines in real GDP. In particular, this is evident in the recessions occurring in the 2000s (2001 and 2008). This is reflected in the lower RMSE for the MIDAS model, which shows an 13% gain over the OLS model.

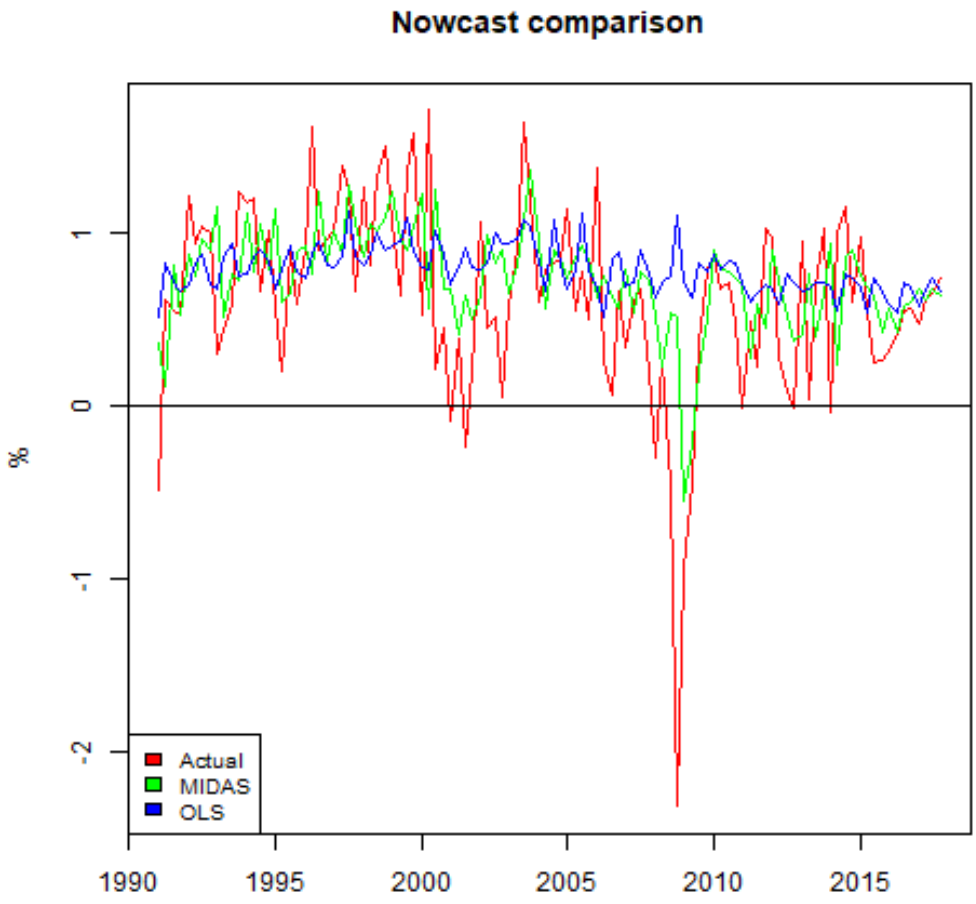

Figure 3: Nowcast using inflation

5 Conclusion

MIDAS models are a useful tool for forecasting low-frequency variables using higher-frequency data. They incorporate the additional information provided by higher-frequency data, unlike models which merely aggregate data to the lowest frequency. However, they also avoid the problems of parameter proliferation, multicollinearity, and low degrees of freedom that would arise from estimating parameters for many high-frequency lags. The MIDAS model does this by calculating weights for the lags using a polynomial function governed by a small number of parameters. Generally, the function suggests decreasing weights as the lags go further back in time. MIDAS models can also be used for now-casting the dependent variable in the current period.

We apply the MIDAS model to economic data, using monthly consumer prices and unemployment rates to forecast quarterly GDP. We seasonally adjust and make stationary the variables to avoid spurious correlation. We compare the MIDAS model forecasts to forecasts using a simple OLS regression where the the high-frequency variables are averaged to the lower frequency. According to the models' root mean squared errors, the MIDAS model outperforms OLS for both CPI and unemployment.

6 Responsibilities for each section

Dina	Introduction and lit review
Alex	Methodology and data description
Doga	Application and R-code

References

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, 158(2):246–261.

Andreou, E., Ghysels, E., and Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, 31(2):240–251.

Armesto, M. T., Engemann, K. M., Owyang, M. T., et al. (2010). Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review*, 92(6):521–36.

Clements, M. P. and Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. *Journal of Business & Economic Statistics*, 26(4):546–554.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. *UCLA, Finance*.

Giles, D. (2017). Explaining the Almon distributed lag model. http://davegiles.blogspot.ca/2017/01/explaining-almon-distributed-lag-model.html.

Gomez-Zamudio, L. M. and Ibarra, R. (2017). Are daily financial data useful for forecasting GDP?: Evidence from Mexico. *Economía*, 17(2):173–203.

Granger IV, C. W., Hyung, N., and Jeon, Y. (2001). Spurious regressions with stationary series. *Applied Economics*, 33(7):899–904.

Marcellino, M. and Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics*, 72(4):518–550.

Monsell, B. C. (2012). X-13-ARIMA-SEATS - a basic seasonal adjustment glossary. https://www.census.gov/srd/www/x13as/glossary.html.

Monsell, B. C., Aston, J. A., and Koopman, S. J. (2003). Toward x-13. *ASA proceedings, business and economic statistics section*.

7 Appendix: R Code

```
############################
### Group project       ###
### Dina Jankovic       ###
### Alexander Lam       ###
### Doga Bilgin         ###
### Topic: Time series  ###
############################


rm(list = ls())
setwd("C:/Users/Helin/Documents/Classes/Carleton/Data Mining I/Group Project")


#### Required libraries, setup, required functions ####
#libraries
if (FALSE){   #Change to TRUE to install packages
  install.packages("seasonal")
  install.packages("midasr")
  install.packages("lubridate")
  install.packages("data.table")
  install.packages("ggplot2")
}
```

```
library(data.table)
library(seasonal)
library(midasr)
library(lubridate)
library(ggplot2)

#functions
rmse <- function(x,y){sqrt(mean((x-y)^2))}

#dates, in quarters
est_st <- c(1980,1)
est_end <- c(1990,4)
rmse_st <-  c(1991,1)
rmse_end <- c(2017,4)

#### Importing Data ####
#initialize lists
data <- NULL
us <- NULL
ca <- NULL

#set up data files for Canada and the US, by frequency
```

```
for (freq in c("q","m","w")){
  data[[freq]] <- data.table(read.csv(
       paste("./Data/",freq,".csv",sep="")
                                  ,na.strings = "#N/A"))
  if (freq=="q") {
    f <- 4
    st <- c(1945,1)
  } else if (freq=="m") {
    f <- 12
    st <- c(1945,1)
  } else if (freq=="w") {
   f <- 365.25/7
   st <- decimal_date(ymd("1945-01-05"))
  } else {
    print("unused frequency")
  }

  data[[freq]] <- data[[freq]][,-1]
  data[[freq]] <- ts(data[[freq]],frequency=f,start=st)

  us[[freq]] <- data[[freq]][,-c(grep("ca",colnames(data[[freq]])))]
  ca[[freq]] <- data[[freq]][,-c(grep("us",colnames(data[[freq]])))]
}
```

```
rm(data,f,st)


#### MIDAS estimation ####
for (var in c("uscpix_nsa","usunemp")){
#variables to use
y <- 100*log(us$q)


#seasonal adjustment
if (var == "uscpix_nsa"){
  x <- 100*log(us$m[,var])
  y <- final(seas(y,x11=""))


  x <-window(final(seas(x,x11=""))
                            ,start=start(x),end=end(x),extend=TRUE)
} else {
  x <- us$m[,c(var)]
}


x_q <- diff(aggregate(x,nfrequency=4,FUN=mean))


y <- diff(y)
x <- diff(x)
```

```
y <- window(y,start=est_st)
x <- window(x,start=est_st)
x_q <- window(x_q,start=est_st)

fcast <- NULL
error <- NULL
i <- 0

#Do RMSE for 1 period ahead over the RMSE period
for (yr in rmse_st[1]:rmse_end[1]){

  if (yr==rmse_end[1]){ #if end quarter is different from 4
    last_qtr = rmse_end[2]
  } else {
    last_qtr = 4
  }

  for (qtr in 1:last_qtr){
    i <- i+1
    date_q = c(yr,qtr-1)
    date_m = c(yr,3*(qtr-1))
    x_est <-window(x,start=est_st,end=date_m)
```

```
    y_est <- window(y,start=est_st,end=date_q)
    midas_temp <- midas_r(y_est~mls(y_est,1,1)+
                            fmls(x_est,2,3,nealmon),
                          start=list(x_est=rep(0,3)))
    x_cur <- window(x,start=date_m+c(0,1),end=date_m+c(0,3))
    f <- forecast(midas_temp,list(x_est=x_cur,method="dynamic"))
    fcast$midas[i] <- f$mean

    x_q_est <- window(x_q,start=est_st,end=date_q)
    midas_temp <- lm(y_est~x_q_est)
    x_q_cur <- window(x_q,start=c(yr,qtr),end=c(yr,qtr))
    fcast$ols[i] <- predict(midas_temp,
                            list(x_q_est=x_q_cur,y_est=last(y_est)))

  }
}

fcast <- as.data.frame(fcast)
fcast <- ts(fcast,start=rmse_st,frequency = 4)
fcast <- ts.intersect(y,fcast) #merge actual (y) with forecasted values

# calculate RMSE
error$midas <- rmse(fcast[,"fcast.midas"], y)
```

```
error$ols <- rmse(fcast[,"fcast.ols"],y)
error
error$midas/error$ols

#plot predicted against forecasted values
png(filename=paste("./Work/nowcast_pred_",var,".png",sep=""))
plot(fcast,plot.type = "s",col=rainbow(n=3,alpha=1),
     ylab="GDP growth, %", xlab="")
legend("bottomleft",legend = c("Actual","MIDAS","OLS")
       , cex=0.8, fill=rainbow(n=3))
title("Nowcast comparison")
abline(h=0)
dev.off()

m <- midas_r(y~mls(y,1,1)+fmls(x,2,3,nealmon),
             start=list(x=rep(0,3)))
ols <- lm(y~x_q)

summary(m)
summary(ols)

}
```