# Fuzzy Cluster Analysis with Mixed Frequency Data

Kaiji Motegi*

July 19, 2014

**Abstract**

This paper develops fuzzy cluster analysis with mixed frequency data. Time series are often sampled at different frequencies like month, quarter, etc. The classic fuzzy cluster analysis simply aggregates all data into the common lowest frequency and then computes a similarity matrix. Such temporal aggregation may yield inaccurate or misleading results due to information loss. Inspired by the growing literature of Mixed Data Sampling (MIDAS) regression technique, this paper proposes a way to construct a similarity matrix exploiting all data available whatever their sampling frequencies are. Empirical illustration using recent Japanese and U.S. macroeconomic indicators suggests that the mixed frequency approach produces clearly different partition trees than the classic low frequency approach does.

## 1 Introduction

Time series are often sampled at different frequencies like month, quarter, year, etc. When the classic fuzzy cluster analysis is applied to multivariate time series data with mixed frequencies, it naïvely aggregates all data into the common lowest frequency and then compute a similarity matrix. A potential problem of this approach is that we are discarding a lot of information on high frequency time series. As a result we may get inaccurate or even misleading implications. It is thus desired to develop a new type of fuzzy cluster analysis that exploits all data available whatever their sampling frequencies are.

Handling mixed frequency data is not an issue limited to fuzzy cluster analysis. This is a universal problem that challenges time series analysis in general. Ghysels, Santa-Clara, and Valkanov (2004), Ghysels, Santa-Clara, and Valkanov (2006), and Andreou, Ghysels, and Kourtellos (2010) propose an innovative regression technique that avoids temporal aggregation: Mixed Data Sampling (MIDAS) regression. Assume each low frequency time period $\tau_L \in \{1, \dots, T_L\}$ contains $m \in \mathbb{N}$ high frequency time periods. The ratio of sampling frequencies, $m$, is equal to 3 for a month vs. quarter mixture, 12 for a month vs. year mixture, and so on. The basic idea of MIDAS regression is to regress a low frequency variable $x_L$ onto all $m$ observations of a high frequency variable $x_H$:

$$x_L(\tau_L) = \alpha + \beta_1 x_H(\tau_L, 1) + \cdots + \beta_m x_H(\tau_L, m) + u_L(\tau_L), \quad \tau_L = 1, \dots, T_L. \tag{1.1}$$

*Faculty of Political Science and Economics, Waseda University. E-mail: motegi@aoni.waseda.jp

$x_H(\tau_L, j)$ is the $j$-th high frequency observation of $x_H$ within low frequency period $\tau_L$. Note that the classic single-frequency regression works on aggregated $x_H$ and hence is written as:

$$x_L(\tau_L) = \alpha + \beta \sum_{j=1}^{m} w_j x_H(\tau_L, j) + u_L(\tau_L), \tag{1.2}$$

where $\boldsymbol{w} = [w_1, \ldots, w_m]'$ represents a linear aggregation scheme that is given, fixed, or pre-determined. It includes flow sampling ($w_j = 1/m$ for $j = 1, \ldots, m$) and stock sampling ($w_m = 1$ and $w_j = 0$ for $j = 1, \ldots, m-1$) as special cases. Clearly, model (1.1) is more general than model (1.2) and hence captures the relationship between $x_H$ and $x_L$ more accurately.

As surveyed by Andreou, Ghysels, and Kourtellos (2011) and Armesto, Engemann, and Owyang (2010), the MIDAS literature is growing very rapidly in the past decade. Most recent development includes Anderson, Deistler, Felsenstein, Funovits, Zadrozny, Eichler, Chen, and Zamani (2012), Ghysels (2012), and McCracken, Owyang, and Sekhposyan (2013). They extend the MIDAS concept to vector autoregression (VAR) in order to treat more than two variables at the same time. Foroni, Ghysels, and Marcellino (2013) provide a survey of mixed frequency VAR models and related literature. Ghysels, Hill, and Motegi (2013) propose Granger causality tests based on Ghysels' (2012) mixed frequency VAR. Ghysels, Hill, and Motegi (2014) invent another mixed frequency Granger causality test that is useful when the ratio of sampling frequencies $m$ is large.

So far the MIDAS framework has never been introduced to fuzzy cluster analysis, and this paper fills that gap. We show via simple Monte Carlo simulations that there exists a certain sort of interdependence between $x_L$ and $x_H$ that the MIDAS regression can capture but the classic low frequency regression cannot. Using the fuzzy cluster analysis with mixed frequency data, we analyze the interdependence between recent Japanese and U.S. macroeconomy. The mixed frequency approach and the conventional low frequency approach produce clearly different partition trees.

The present paper is organized as follows. Section 2 describes our methodology. Section 3 runs the Monte Carlo simulations. Section 4 presents the empirical application. Section 5 concludes the paper.

## 2 Methodology

Although our methodology could be applied to an arbitrary number of sampling frequencies, this paper assumes for expositional simplicity that there are only two: either high frequency or low frequency. Suppose that we have $K_H$ high frequency variables $x_{H,1}, \ldots, x_{H,K_H}$ and $K_L$ low frequency variables $x_{L,1}, \ldots, x_{L,K_L}$. We thus have $K = K_H + K_L$ variables in total. Each low frequency time period $\tau_L$ have $m$ high frequency periods. The ratio of sampling frequencies $m$ may depend on low frequency time periods in some applications like week vs. month (one month contains four or five weeks). This paper assumes for simplicity that $m$ is constant over time (e.g. month vs. quarter where $m$ is always 3).

Consider a low frequency time period $\tau_L$. In the first high frequency period within $\tau_L$, we observe $x_{H,1}(\tau_L, 1), \ldots, x_{H,K_H}(\tau_L, 1)$. In the second high frequency period within $\tau_L$, we observe $x_{H,1}(\tau_L, 2)$, $\ldots$, $x_{H,K_H}(\tau_L, 2)$, and so on. In the last high frequency period within $\tau_L$, we observe $x_{H,1}(\tau_L, m)$,

$\ldots, x_{H,K_H}(\tau_L, m)$ as well as $x_{L,1}(\tau_L), \ldots, x_{L,K_L}(\tau_L)$. The assumption that $x_L$ is observed at the last high frequency period is just by convention, and can be relaxed if desired. See Figure 1 for a visual explanation of these notations. In the figure there are assumed to be only one high frequency variable $x_H$ and only one low frequency variable $x_L$ (i.e. $K_H = K_L = 1$). We sequentially observe $x_H(\tau_L, 1)$, $x_H(\tau_L, 2), \ldots, x_H(\tau_L, m)$, $x_L(\tau_L)$ in low frequency period $\tau_L$.

Before describing our own methodology, let us recall how we usually apply the classic fuzzy cluster analysis to mixed frequency data. What we used to do is aggregating each high frequency variable into low frequency first of all: $x_{H,k}(\tau_L) = \sum_{j=1}^{m} w_j^k x_{H,k}(\tau_L, j)$ for $k = 1, \ldots, K_H$. $\boldsymbol{w}^k = [w_1^k, \ldots, w_m^k]'$ represents a linear aggregation scheme for $x_{H,k}$ (e.g. stock sampling, flow sampling, etc.). Now we have all $K$ variables having a single frequency, so we compute a similarity matrix in a usual way. A well-known similarity measure is correlation coefficient or something similar. Consider $x_{L,1}$ and aggregated $x_{H,1}$ for instance. A common way of defining a similarity measure between these two variables is to run ordinary least squares (OLS) with respect to a linear regression model:

$$x_{L,1}(\tau_L) = \alpha + \beta x_{H,1}(\tau_L) + u_{L,1}(\tau_L), \quad \tau_L = 1, \ldots, T_L \tag{2.3}$$

and then calculate $R^2$. Let $\hat{\alpha}$ and $\hat{\beta}$ be OLS estimators for $\alpha$ and $\beta$, respectively. Using $\hat{\alpha}$ and $\hat{\beta}$, we compute residuals $\hat{u}_{L,1}(\tau_L) = x_{L,1}(\tau_L) - \hat{\alpha} - \hat{\beta} x_{H,1}(\tau_L)$. $R^2$ is defined as follows:

$$R^2 = 1 - \frac{\sum_{\tau_L=1}^{T_L} \hat{u}_{L,1}^2(\tau_L)}{\sum_{\tau_L=1}^{T_L} (x_{L,1}(\tau_L) - \bar{x}_{L,1})^2},$$

where $\bar{x}_{L,1} = (1/T_L) \sum_{\tau_L=1}^{T_L} x_{L,1}(\tau_L)$. It is well-known that $R^2$ is equal to the squared correlation coefficient between $x_{L,1}$ and aggregated $x_{H,1}$ in this bivariate setting.

When the classic fuzzy cluster analysis computes $R^2$ between two high frequency variables (say $x_{H,1}$ and $x_{H,2}$), past papers often work on aggregated $x_{H,1}$ and aggregated $x_{H,2}$. They usually create a single-frequency setting at the very beginning of analysis, so never exploit original high frequency observations of $x_{H,1}, \ldots, x_{H,K_H}$. After getting $R^2$ for all possible pairs, the usual clustering procedure (e.g. Zadeh's method, Ward's method) used to be applied in order to draw a partition tree. Finally, an optimal level of the partition tree is determined by fuzzy theory.

A potential problem of these classic procedures is that the temporal aggregation of high frequency variables may cause inaccurate or misleading results due to information loss. Assuming $m = 3$ for example, the existing approach discards roughly two-thirds of the entire information contained in original high frequency variables. Information loss gets even larger as $m$ increases, a typical example being month vs. year with $m = 12$.

Now we explain how to exploit mixed frequency data efficiently. Based on the MIDAS literature, it is straightforward to generalize model (2.3) to a mixed frequency framework:

$$x_{L,1}(\tau_L) = \alpha + \beta_1 x_{H,1}(\tau_L, 1) + \cdots + \beta_m x_{H,1}(\tau_L, m) + u_{L,1}(\tau_L), \quad \tau_L = 1, \ldots, T_L. \tag{2.4}$$

We run OLS with respect to model (2.4) and then compute $R^2$, or adjusted $R^2$ if we want to take model

parsimony into account. Since model (2.4) has $m+1$ regressors, adjusted $R^2$ is computed as follows:

$$\bar{R}^2 = 1 - \frac{\frac{1}{T_L-(m+1)}\sum_{\tau_L=1}^{T_L}\hat{u}_{L,1}^2(\tau_L)}{\frac{1}{T_L-1}\sum_{\tau_L=1}^{T_L}(x_{L,1}(\tau_L)-\bar{x}_{L,1})^2}.$$

The MIDAS framework does not involve temporal aggregation, so it allows us to work on high frequency data when we compute (adjusted) $R^2$ between two high frequency variables, say $x_{H,1}$ and $x_{H,2}$. Simply regress one onto the other in a single-frequency (but high frequency) setting, and then calculate (adjusted) $R^2$. After getting (adjusted) $R^2$ for all possible pairs, just a usual clustering procedure is applied (e.g. drawing a partition tree, finding an optimal level of the tree based on fuzzy theory, etc.).

# 3 Illustrative Simulation Study

We run simple Monte Carlo experiments in order to highlight an advantage of the MIDAS regression approach over the traditional low frequency approach. We simulate 100,000 samples from a linear data generating process (DGP):

$$x_L(\tau_L) = 0.2x_H(\tau_L,1) + 0.1x_H(\tau_L,2) - 0.2x_H(\tau_L,3) + \epsilon_L(\tau_L), \quad \tau_L = 1,\ldots,10. \tag{3.5}$$

$x_H(\tau_L,1), x_H(\tau_L,2), x_H(\tau_L,3)$ are mutually and serially uncorrelated standard normal random numbers. $\epsilon_L(\tau_L)$ are serially uncorrelated random numbers drawn from $N(0,0.1)$. We assume independence between $x_H$ and $\epsilon_L$. The ratio of sampling frequencies, $m$, is set to be 3 so that this experiment can be thought of as a month vs. quarter analysis just like Section 4. Sample size in terms of low frequency is assumed to be only 10 quarters in order to match the empirical application below. Increasing the sample size would not change the main conclusion of this experiment, however.

In the true DGP $x_H$ does have a relevant impact on $x_L$, but we have both positive and negative impacts at the same time. $x_H(\tau_L,1)$ and $x_H(\tau_L,2)$ have positive coefficients (0.2 and 0.1), while $x_H(\tau_L,3)$ has a negative coefficient of -0.2. It is not uncommon to encounter mixed signs in theory and practice of economics.

For each sample we fit a MIDAS regression model:

$$x_L(\tau_L) = \alpha + \beta_1 x_H(\tau_L,1) + \beta_2 x_H(\tau_L,2) + \beta_3 x_H(\tau_L,3) + u_L(\tau_L) \tag{3.6}$$

as well as the classic low frequency regression model:

$$x_L(\tau_L) = \alpha + \beta x_H(\tau_L) + u_L(\tau_L),$$

where we assume flow sampling $x_H(\tau_L) = (1/3)\sum_{j=1}^{3} x_H(\tau_L,j)$. For each regression model we compute adjusted $R^2$ and then plot a histogram in order to compare the model adequacy.

Figure 2 plots the histograms of adjusted $R^2$, written as $\bar{R}^2$. Panel (a) is concerned with the MIDAS regression, while Panel (b) is concerned with the low frequency regression. The horizontal axis has $\bar{R}^2$,

while the vertical axis has the normalized frequency that adds up to 1.

In Panel (a), about 65% of the total replications get $\bar{R}^2$ beyond 0.9, and about 90% of the total replications get $\bar{R}^2$ beyond 0.8. This means that the MIDAS regression model fits simulated data very well, an expected result since model (3.6) is correctly specified relative to DGP (3.5).

In Panel (b), about 55% of the total replications get negative $\bar{R}^2$, and about 70% of the total replications get $\bar{R}^2$ below 0.1. This means that the classic low frequency regression model with flow sampling cannot capture the underlying relationship between $x_L$ and $x_H$ at all. A key here is that we have both positive coefficients (i.e. 0.1, 0.2) and a negative coefficient (i.e. -0.2) in the DGP. Flow aggregation takes a arithmetic mean of $x_H(\tau_L, 1), x_H(\tau_L, 2), x_H(\tau_L, 3)$ and hence the positive impact and negative impact offset each other, yielding spuriously weak impact of aggregated $x_H$ on $x_L$. This example highlights an advantage of the MIDAS regression approach which is free of temporal aggregation.

# 4 Empirical Application

Using the mixed frequency fuzzy cluster analysis, we investigate the interaction among recent macroeconomic time series in Japan and the U.S. Section 4.1 describes data while Section 4.2 presents empirical findings.

## 4.1 Data

For each of Japan and the U.S., we prepare monthly unemployment rate (UR), monthly consumer price index (CPI), and quarterly real gross domestic product (GDP). All data are publicly available online. Japanese unemployment and CPI can be found at the website of Statistics Bureau, the Ministry of Internal Affairs and Communications. Japanese GDP can be found at the website of Cabinet Office. All U.S. series are downloadable at Federal Reserve Economic Data (FRED).

Note that unemployment rate and CPI are released each month while GDP is released each quarter. This is a typical example where the mixed frequency approach matters. We have four high frequency variables ($K_H = 4$) and two low frequency variables ($K_L = 2$). We take year-to-year change in monthly unemployment rate to remove potential seasonal effects. Similarly, we take year-to-year growth rate of monthly CPI and quarterly GDP.

Unemployment, CPI, and GDP are generally regarded as key indicators representing overall macroeconomic performance. In particular, negative correlation between unemployment rate and CPI is known as the Phillips Curve. Also, negative correlation between unemployment rate and GDP is known as the Okun Curve.

Our sample period covers July 2011 through March 2014, which has 30 months (or 10 quarters). This is a relatively small sample, and the largest possible sample we could take is January 1981 - March 2014 (Japan's quarterly real GDP in 1980 or before cannot be retrieved). Such a large sample would most likely contain many structural breaks, however. We have a number of historical events that have most likely changed the interdependence structure among Japanese and U.S. macroeconomic time series. To name a few, we had Japan's stock market bubble and burst in late 1980s, dot-com bubble in late 1990s,

subprime mortgage crisis in late 2000s, and a devastating earthquake in Japan in March 2011. Our sample July 2011 - March 2014 does not contain any of them, and can be thought of as a relatively stable sample period. A virtue of the mixed frequency approach is that it allows us to work on such a short sample period. If we took the classic low frequency approach, the number of observations would be only 10 for each series. Since we are taking the mixed frequency approach, the number of observations is 30 for unemployment and CPI and 10 for GDP.

Figure 3 plots year-to-year change in monthly unemployment rate, year-to-year growth rate of monthly CPI, and year-to-year growth rate of quarterly real GDP in Japan and the U.S. Panels (a)-(d) plot the monthly series, while Panels (e) and (f) plot the quarterly series. Vertical axes for Panels (a)-(c) span $[-1.5, 2]$, while the vertical axes for Panels (d)-(f) span $[-1, 5]$. Sample period covers July 2011 through March 2014.

Panels (a) and (b) indicate that unemployment rate is declining slowly but consistently in both Japan and the U.S. Panel (f) agrees with Panel (b) that the U.S. economy is expanding at a steady rate, showing the real GDP growth between 1% and 3.5%. Panel (e), however, suggests that Japan ran into a short recession late 2012 and early 2013. The real GDP growth in that period is marginally below zero. The discrepancy between Panels (a) and (e) implies that unemployment and GDP measure different aspects of macroeconomic activity, at least in recent Japan.

Panels (c) and (d) highlight the difference between Japanese goods market and U.S. goods market. Japan had been suffering from prolonged deflation since 1990s although that seems to be over very recently. As seen in Panel (c), Japan's inflation got positive and began rising in the middle of 2013, reaching about 1.5% in 2014. The U.S. in contrast has never experienced deflation in the past decades. Panel (d) shows moderately high U.S. inflation (2-4%) until middle 2012 and stable inflation (1-2%) since then.

Table 1 reports sample mean, median, minimum, maximum, standard deviation, skewness, and kurtosis of each series plotted in Figure 3. Table 1 provides basically the same implications as Figure 3. First, unemployment rates in Japan and the U.S. are declining on average at a slow rate. The mean is -0.36 for Japan and -0.80 for the U.S., while the standard deviation is 0.20 for Japan and 0.23 for the U.S. Second, Japan is moving from deflation to inflation while the U.S. is having consistent inflation. The minimum is -0.90 for Japan and 0.92 for the U.S. while the maximum is 1.61 for Japan and 3.85 for the U.S. Third, the U.S. has higher and more stable real GDP growth than Japan in our sample period. The mean is 1.32 for Japan and 2.20 for the U.S. while the standard deviation is 1.50 for Japan and 0.65 for the U.S.

### 4.2 Empirical Results

We apply the mixed frequency fuzzy cluster analysis to the six variables described above. Recall that (1) adjusted $R^2$, written as $\bar{R}^2$, between a monthly variable and a quarterly variable is computed via the MIDAS regression (2.4), (2) $\bar{R}^2$ between a monthly variable and another monthly variable is computed on the standard single-frequency (but monthly) basis, and (3) $\bar{R}^2$ between a quarterly variable and the other quarterly variable is computed on the standard single-frequency (quarterly) basis.

For comparison, we also implement the classic fuzzy cluster analysis based on *aggregated* high frequency variables (cfr. model (2.3)). Since change in unemployment rate and the growth rate of CPI are both flow variables, we employ flow aggregation $x_H(\tau_L) = (1/3)\sum_{j=1}^{3} x_H(\tau_L, j)$ for each high frequency series.

We employ Zadeh's method (a.k.a. nearest neighbor method) and Ward's method to draw partition trees. We are interested in whether the mixed frequency approach and the classic low frequency approach produce different partition trees, and how they are different if any.

Figure 4 plots partition trees based on Zadeh's method. Panel (a) is concerned with mixed frequency approach which works on monthly unemployment rate, monthly CPI, and quarterly GDP. Panel (b) is concerned with the classic low frequency approach which works on quarterly unemployment rate, quarterly CPI, and quarterly GDP. Similarity value (i.e. adjusted $R^2$) is put for each level.

Evidently, the mixed frequency approach and the low frequency approach produce different partition trees. In the mixed frequency case U.S. unemployment and U.S. GDP merge first, which corresponds to the Okun's law (i.e. negative correlation between unemployment and GDP). In the low frequency case Japanese CPI and Japanese GDP merge first and then the U.S. Okun-law relation shows up. This suggests that the mixed frequency approach emphasizes the U.S. Okun-law relation more than the low frequency approach does.

There is another difference between Panels (a) and (b) when there are three clusters. Both partition trees have the same three clusters: (1) U.S. unemployment, U.S. GDP, and Japanese unemployment, (2) Japanese CPI and Japanese GDP, and (3) U.S. CPI. How they merge differs across the trees, however. In the mixed frequency case (1) and (2) merge and then (3) joins them. In the low frequency case (1) and (3) merge and then (2) joins them.

We now determine an optimal level of each partition tree. There are three common ways to calculate cluster size at each level: Max approach, Power mean approach, and Arithmetic mean approach (see Chapter 2 of Yamashita and Takizawa (2010) for details). For each approach we find the optimal level and put a letter "M", "P", or "A" in Figure 4. All approaches agree that the optimal level in Panel (a) is 0.19, where we have (1) U.S. unemployment, U.S. GDP, and Japanese unemployment, (2) Japanese CPI and Japanese GDP, and (3) U.S. CPI. All approaches agree that the optimal level in Panel (b) is 0.27, where we have exactly same clusters (1), (2), and (3). Therefore, fuzzy decision suggests that the two partition trees are similar at least at the optimal level. In this sense taking the mixed frequency approach does not necessarily change an essential part of a partition tree, although it does change how individual variables reach the optimal level and how optimal clusters merge each other.

Figure 5 plots partition trees based on Ward's method. Panel (a) is concerned with mixed frequency approach, while Panel (b) is concerned with the classic low frequency approach. Similarity value (i.e. standardized adjusted $R^2$) is put for each level.

As in Figure 4, the mixed frequency approach emphasizes the U.S. Okun-law relation more than the low frequency approach does. In the mixed frequency case U.S. unemployment and U.S. GDP merge first, as seen in Panel (a) of Figure 5. In the low frequency case Japanese CPI and Japanese GDP merge first and then the U.S. Okun-law relation shows up.

Further, there is another difference when there are three clusters. Both partition trees have the same

three clusters: (1) U.S. unemployment and U.S. GDP, (2) Japanese unemployment and U.S. CPI, and (3) Japanese CPI and Japanese GDP. How they merge differs across the trees, however. In the mixed frequency case (2) and (3) merge and then (1) joins them. In the low frequency case (1) and (3) merge and then (2) joins them.

We now determine an optimal level of each partition tree with respect to Ward's method. For Panel (a), fuzzy decision with max approach chooses 0.66, where we have (i) U.S. unemployment and U.S. GDP, (ii) Japanese unemployment, (iii) U.S. CPI, (iv) Japanese CPI, and (v) Japanese GDP. Fuzzy decision with power mean approach and arithmetic mean approach agree with each other that the optimal level is 0.34, where we have (1) U.S. unemployment and U.S. GDP, (2) Japanese unemployment and U.S. CPI, and (3) Japanese CPI and Japanese GDP. For Panel (b), fuzzy decision with max approach chooses 0.52, where we have (i') Japanese CPI and Japanese GDP, (ii') U.S. unemployment, (iii') U.S. GDP, (iv') Japanese unemployment, and (v') U.S. CPI. Fuzzy decision with power mean approach and arithmetic mean approach agree with each other that the optimal level is 0.31, where we have exactly (1), (2), and (3). Hence, we reach the same optimal clusters across Panels (a) and (b) if we take the power mean approach or arithmetic mean approach. This result again suggests that taking the mixed frequency approach does not necessarily change a core part of a partition tree, although it does change how individual variables reach the optimal level and how optimal clusters merge each other.

In summary, Figures 4 and 5 suggest that taking the mixed frequency approach instead of the low frequency approach may change empirical implications, whether Zadeh's method or Ward's method is used. Optimal levels determined by fuzzy theory are likely unchanged, but the detailed structure of a partition tree does change significantly.

# 5 Conclusions

Time series are often sampled at different frequencies like month, quarter, etc. When the classic fuzzy cluster analysis is applied to multivariate time series data with mixed frequencies, it naïvely aggregates all data into the common lowest frequency and then compute a similarity matrix. A potential problem of this approach is that we are discarding a lot of information on high frequency time series. As a result we may get inaccurate or even misleading implications.

To resolve this issue, we have proposed a new type of fuzzy cluster analysis that exploits all data available whatever their sampling frequencies are. We use the Mixed Data Sampling (MIDAS) regression technique that is increasingly popular in recent time series econometrics. Assuming each low frequency period $\tau_L$ contains $m$ high frequency periods, the MIDAS regression model regresses a low frequency variable $x_L$ onto all $m$ observations of a high frequency variable $x_H$. We compute (adjusted) $R^2$ from the MIDAS regression and then construct a similarity matrix just as usual.

We show via simple Monte Carlo simulations that the mixed frequency approach better captures the underlying relationship between $x_L$ and $x_H$ than the existing low frequency approach. In particular, the mixed frequency approach matters when the high frequency observations of $x_H$ have positive and negative impacts on $x_L$ at the same time.

We study recent Japan-U.S. macroeconomy, comparing the new fuzzy cluster analysis associated

with the MIDAS regression and the classic fuzzy cluster analysis that works on aggregated single-frequency data. The former works on monthly unemployment, monthly inflation, and quarterly GDP, while the latter works on quarterly unemployment, quarterly inflation, and quarterly GDP. It turns out that the mixed frequency approach and the low frequency approach produce clearly different partition trees, whether we use Zadeh's method (a.k.a. nearest neighbor method) or Ward's method. In particular, correlation between U.S. unemployment and U.S. GDP (i.e. the Okun law) is more emphasized in the mixed frequency case. Optimal levels determined by fuzzy theory are likely unchanged, but the detailed structure of a partition tree does change by switching from the low frequency approach to the mixed frequency approach.

## References


ANDERSON, B. D. O., M. DEISTLER, E. FELSENSTEIN, B. FUNOVITS, P. ZADROZNY, M. EICHLER, W. CHEN, AND M. ZAMANI (2012): "Identifiability of Regular and Singular Multivariate Autoregressive Models from Mixed Frequency Data," in *51st Conference on Decision and Control*, pp. 184–189, Maui, HI. IEEE Control Systems Society.

ANDREOU, E., E. GHYSELS, AND A. KOURTELLOS (2010): "Regression Models with Mixed Sampling Frequencies," *Journal of Econometrics*, 158, 246–261.

——— (2011): "Forecasting with Mixed-Frequency Data," in *Oxford Handbook of Economic Forecasting*, ed. by M. Clements, and D. Hendry, pp. 225–245.

ARMESTO, M., K. ENGEMANN, AND M. OWYANG (2010): "Forecasting with Mixed Frequencies," *Federal Reserve Bank of St. Louis Review*, 92, 521–536.

FORONI, C., E. GHYSELS, AND M. MARCELLINO (2013): "Mixed Frequency Approaches for Vector Autoregressions," in *VAR Models in Macroeconomics, Financial Econometrics, and Forecasting - Advances in Econometrics*, ed. by T. Fomby, and L. Killian, vol. 31.

GHYSELS, E. (2012): "Macroeconomics and the Reality of Mixed Frequency Data," Working paper, University of North Carolina at Chapel Hill.

GHYSELS, E., J. B. HILL, AND K. MOTEGI (2013): "Testing for Granger Causality with Mixed Frequency Data," Working paper, University of North Carolina at Chapel Hill.

——— (2014): "Regression-Based Mixed Frequency Granger Causality Tests," Working paper, University of North Carolina at Chapel Hill.

GHYSELS, E., P. SANTA-CLARA, AND R. VALKANOV (2004): "The MIDAS Touch: Mixed Data Sampling Regression Models," Working Paper, UCLA and UNC.

——— (2006): "Predicting volatility: Getting the Most out of Return Data Sampled at Different Frequencies," *Journal of Econometrics*, 131, 59–95.

MCCRACKEN, M., M. OWYANG, AND T. SEKHPOSYAN (2013): "Real-Time Forecasting with a Large Bayesian Block Model," Discussion Paper, Federal Reserve Bank of St. Louis and Bank of Canada.

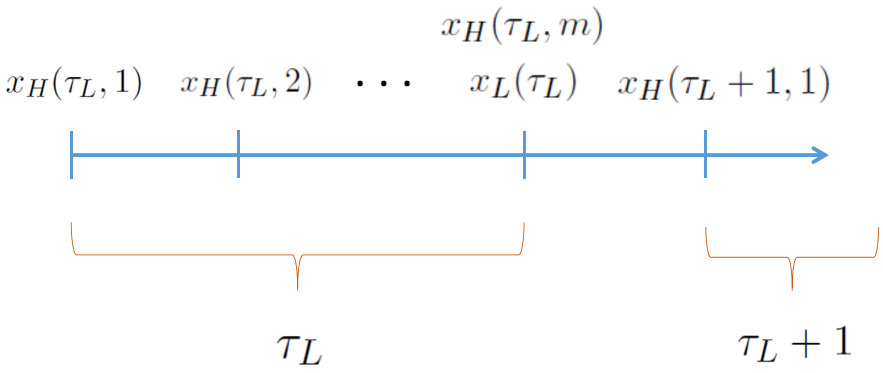
YAMASHITA, H., AND T. TAKIZAWA (2010): *Fuzzy Theory*. Tokyo: Kyoritsu Shuppan Co., Ltd., in Japanese.

## Tables and Figures

Table 1: Sample Statistics

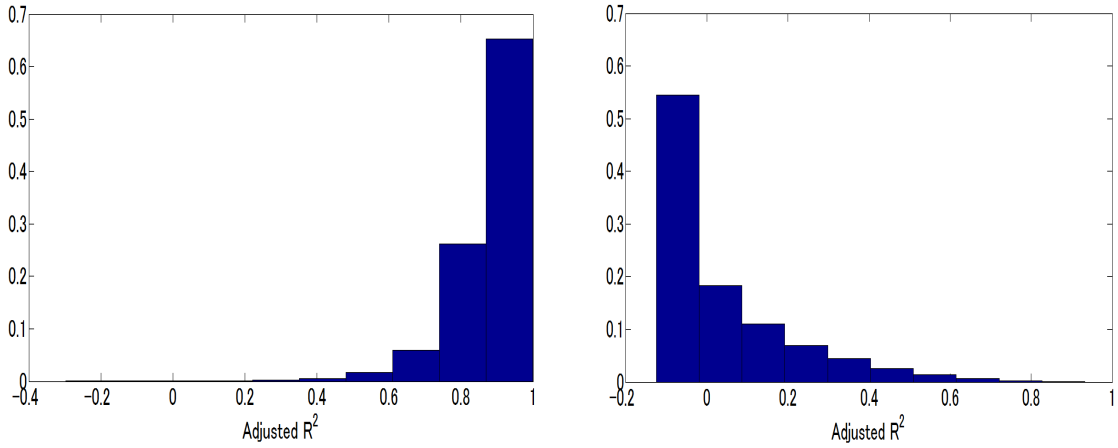
Note: This table reports sample mean, median, minimum, maximum, standard deviation, skewness, and kurtosis of each series from July 2011 through March 2014, which has 30 months (or 10 quarters). We have three series for each of Japan and the U.S.: year-to-year change in monthly unemployment rate, year-to-year growth rate of monthly consumer price index, and year-to-year growth rate of quarterly gross domestic product.

| | Mean | Median | Min. | Max. | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Monthly UR (JP) | -0.36 | -0.30 | -0.90 | 0.10 | 0.20 | -0.50 | 3.93 |
| Monthly UR (US) | -0.80 | -0.80 | -1.30 | -0.30 | 0.23 | -0.28 | 3.18 |
| Monthly CPI (JP) | 0.24 | 0.10 | -0.90 | 1.61 | 0.74 | 0.65 | 2.33 |
| Monthly CPI (US) | 2.06 | 1.76 | 0.92 | 3.85 | 0.85 | 0.89 | 2.60 |
| Quarterly GDP (JP) | 1.32 | 1.31 | -0.48 | 3.22 | 1.50 | 0.04 | 0.94 |
| Quarterly GDP (US) | 2.20 | 2.01 | 1.32 | 3.27 | 0.65 | 0.46 | 2.03 |




Note: This figure explains a standard notation in the Mixed Data Sampling (MIDAS) literature. Assume there are only one high frequency variable $x_H$ and only one low frequency variable $x_L$. In low frequency period $\tau_L$, we sequentially observe $x_H(\tau_L, 1), x_H(\tau_L, 2), \ldots, x_H(\tau_L, m), x_L(\tau_L)$.

Figure 1: Visual Explanation of Mixed Frequency Time Series

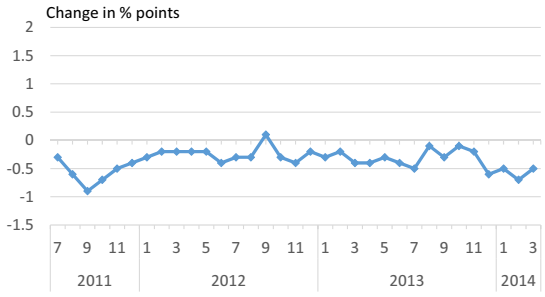



(a) Mixed Frequency Approach

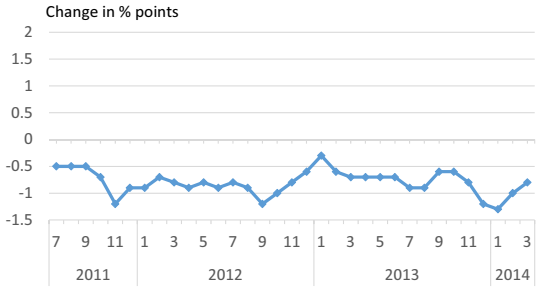
(b) Low Frequency Approach

Note: This figure plots the histograms of adjusted $R^2$ computed through Monte Carlo simulations. Panel (a) is concerned with the MIDAS regression that regresses a low frequency variable $x_L$ onto high frequency observations of $x_H$, while Panel (b) is concerned with the conventional low frequency regression that regresses $x_L$ onto *aggregated* high frequency variable $x_H$. The horizontal axis has adjusted $R^2$, while the vertical axis has the normalized frequency that adds up to 1.

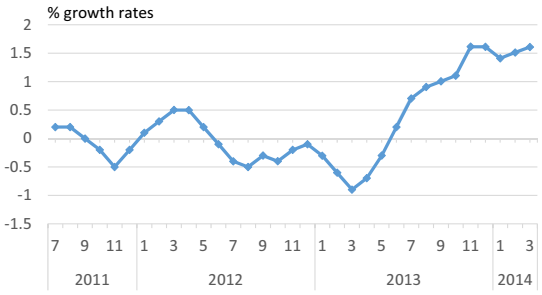
Figure 2: Histograms of Adjusted $R^2$ (Monte Carlo Simulations)



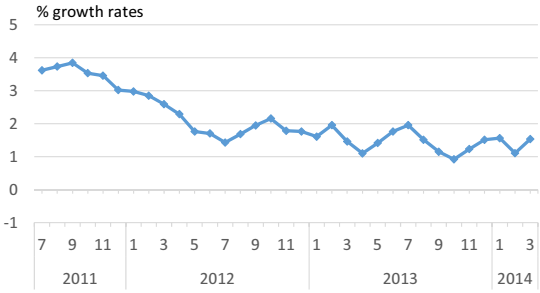

(a) Monthly Unemployment Rate (Japan)



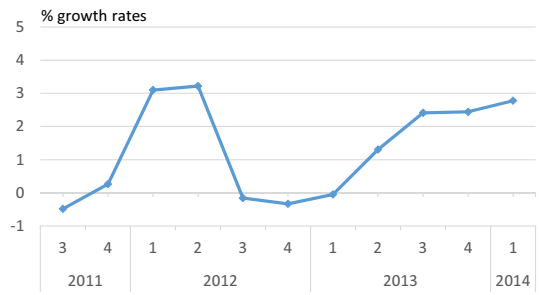

(b) Monthly Unemployment Rate (US)



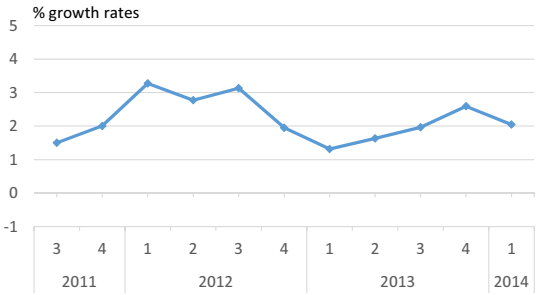

(c) Monthly CPI (Japan)




(d) Monthly CPI (US)



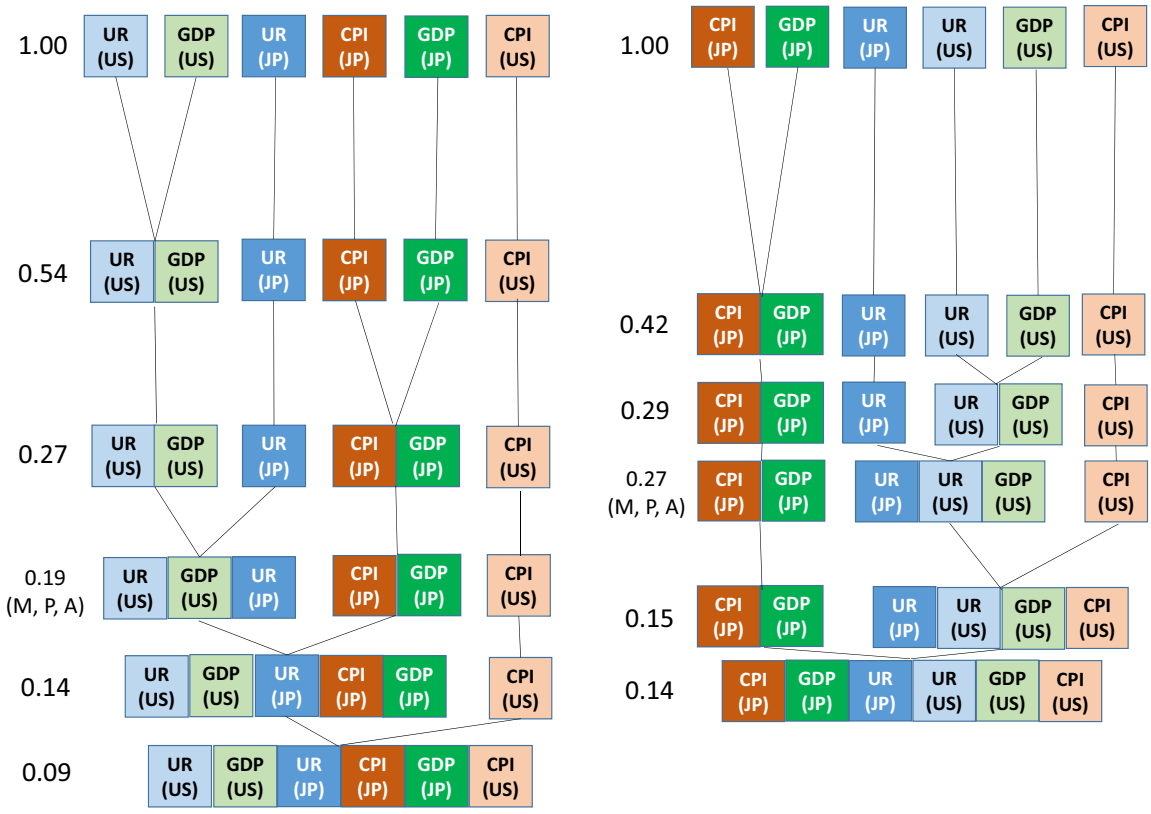

(e) Quarterly GDP (Japan)




(f) Quarterly GDP (US)

Note: This figure plots year-to-year change in monthly unemployment rate, year-to-year growth rate of monthly consumer price index, and year-to-year growth rate of quarterly real gross domestic product in Japan and the U.S. Panels (a)-(d) plot the monthly series, while Panels (e) and (f) plot the quarterly series. Vertical axes for Panels (a)-(c) span $[-1.5, 2]$, while the vertical axes for Panels (d)-(f) span $[-1, 5]$. Sample period covers July 2011 through March 2014, which has 30 months (or 10 quarters).

Figure 3: Monthly Unemployment Rate, Monthly CPI, and Quarterly GDP

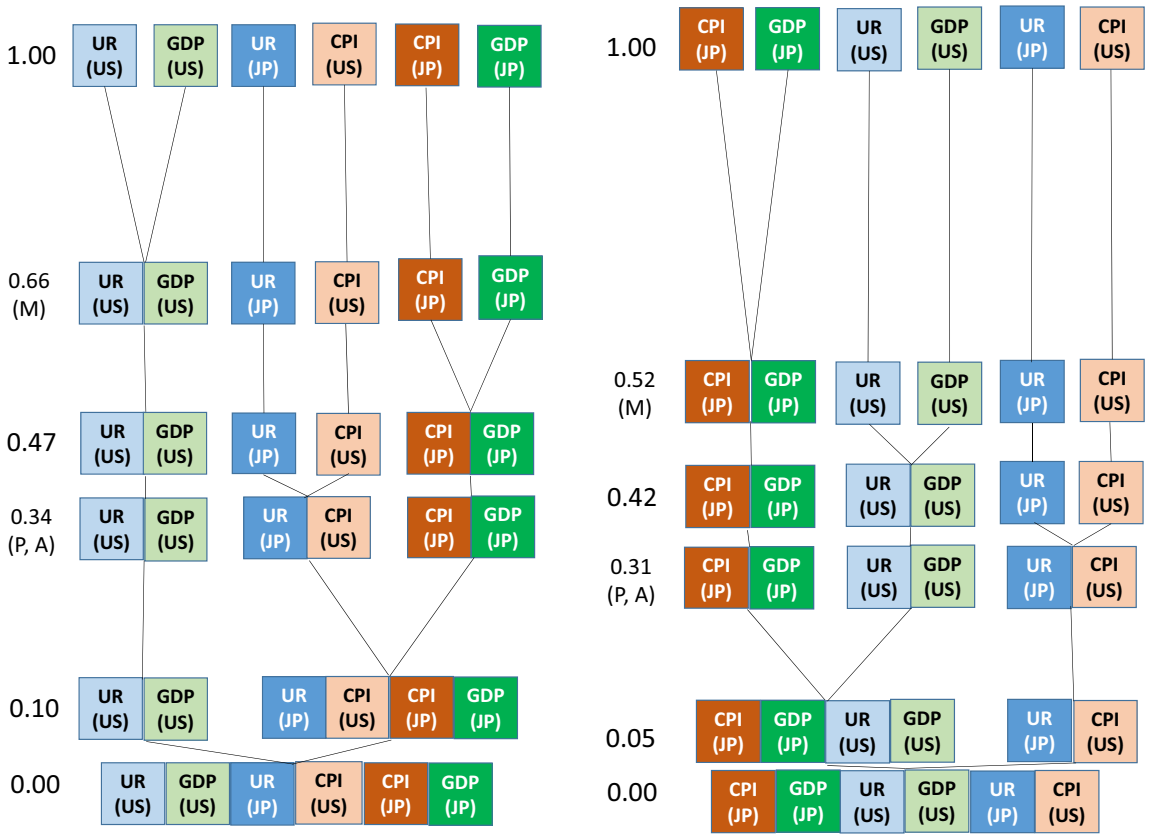



(a) Mixed Frequency Approach

(b) Low Frequency Approach

Note: This figure plots partition trees based on Zadeh's method (a.k.a. nearest neighbor method). Panel (a) is concerned with mixed frequency approach which works on monthly unemployment rate, monthly consumer price index, and quarterly gross domestic product. Panel (b) is concerned with the classic low frequency approach which works on quarterly unemployment rate, quarterly CPI, and quarterly GDP. Similarity value (i.e. adjusted $R^2$) is put for each level. There are three common ways to calculate cluster size at each level: Max approach, Power mean approach, and Arithmetic mean approach (see Chapter 2 of Yamashita and Takizawa (2010) for details). For each approach we find the optimal level using fuzzy decision theory and put a letter "M", "P", or "A".

Figure 4: Partition Trees (Zadeh's Method)




(a) Mixed Frequency Approach

(b) Low Frequency Approach

Note: This figure plots partition trees based on Ward's method. Panel (a) is concerned with mixed frequency approach which works on monthly unemployment rate, monthly consumer price index, and quarterly gross domestic product. Panel (b) is concerned with the classic low frequency approach which works on quarterly unemployment rate, quarterly CPI, and quarterly GDP. Similarity value (i.e. standardized adjusted $R^2$) is put for each level. There are three common ways to calculate cluster size at each level: Max approach, Power mean approach, and Arithmetic mean approach (see Chapter 2 of Yamashita and Takizawa (2010) for details). For each approach we find the optimal level using fuzzy decision theory and put a letter "M", "P", or "A".

Figure 5: Partition Trees (Ward's Method)