Erasmus University Rotterdam

Bachelor Thesis

Data Aggregation in MIDAS Models: Improving Forecasting through Optimal Data Piling

Author:
Scipio Postmes (386125)

Supervisor:
Prof. dr. P.H.B.F. Franses

Second Assessor:
Prof. dr. R. Paap

A thesis submitted in fulfillment of the requirements
for the degree of
the International Bachelor Econometrics and Operations Research

in the direction of Business Analytics & Quantitative Marketing

Erasmus School of Economics

2nd July 2017

ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

Abstract

The MIxed DAta Sampling (MIDAS) model has proven to be a valuable tool in the modeling of data sampled at different frequencies. With this new possibility arises the option to aggregate explanatory data into intermediate frequencies before regressing upon them. As such the room for noise can be decreased. This paper studies the added value of data piling in the use of MIDAS models. It finds that oftentimes there is an intermediate frequency of data aggregation that produces better forecasts than either the raw data or the fully aggregated data. Furthermore it appears that R^2 and Akaike Information Criteria are in some cases accurate predictors of model optimality. However, the number of cases in which they are wrong remains too large. Still data piling should be regarded as a valuable addition to today's econometric Time Series toolbox.

Contents

Chapter 1

Introduction

In recent years, much has been written about the MIxed DAta Sampling (MIDAS) models that have been introduced by Ghysels et al. in 2004 (Ghysels et al., 2004). Following the demonstrated use in financial applications in the context of volatility forecasting by Ghysels et al. (2006) were many others, including Clements and Galvão (2008), who show that the MIDAS model also has value in a macroeconomic context and add an autoregressive component. MIDAS models were also recently used by Andreou et al. (2010) and Monteforte and Moretti (2010) to forecast quarterly GDP using daily observed financial variables. Franses (2016) analyses different specifications of the MIDAS model and evaluates their performance in both simulated and empirical context. The ability to model data sampled at a low frequency using data sampled at a higher frequency opens doors to many new possibilities.

But besides the various areas of application and broad range of specifications there are also some entirely new possibilities for the use of MIDAS models. This research will focus on the researching optimal levels of data piling, which is possible due to the existence and functionality of MIDAS models. For instance, if one analyses daily financial data in order to forecast quarterly or yearly GDP, the case might be made that not every day has its 'own' unique effect, and trying to forecast using all days might leave too much room for unnecessary noise. It might make more sense to aggregate the daily observations into weekly, monthly or even quarterly observations. This research will attempt to construct a technique to find the optimal data aggregation level in order to achieve the best forecasts.

Chapter 2

Experiment Description

This section describes the models and techniques used in order to investigate the efficiency of different kinds of MIDAS models. It first discusses Unrestricted MIDAS (U-MIDAS) models and then addresses the Almon lag and its use in MIDAS models.

2.1 Defining Optimal Forecasting Performance

Since a study is conducted on the optimal level of explanatory variable aggregation, we need to define optimality. The optimal model will be defined as the model that produces the best forecasts in terms of Root Mean Squared Prediction Error (RMSPE) performance. The study of RMSPE performance is interesting as it includes punishments for both bias and inefficiency. The goal of this analysis is to find a way to estimate the data aggregation level that produces the best forecasts based on the parameter estimates of a hold-out sample. That is, based on test values originating from the estimation sample we need to be able to identify what level of data aggregation will deliver the best forecasting performance.

Given that the MIDAS structures properly estimate the parameters, there arises a new set of possibilities in constructing forecasts for the low-frequency dependent variable. For example, say we take annual unemployment rates as our dependent variable, and weekly observations of the number of employees of a staffing agency as our independent variables. It is now possible to apply a MIDAS structure to these data and, as such, construct a one-step-ahead forecast based on our weekly data. However, it would also be possible to aggregate the weekly observations into monthly ones and, consequently, use those

monthly observations to construct a (monthly-to-yearly) MIDAS model. Similarly, we could construct quarterly or biannual observations and corresponding MIDAS structures.

2.2 Unrestricted MIDAS Models

The Unrestricted MIxed DAta Sampling (U-MIDAS) model was proposed by Koenig et al. (2003) and was also considered by Clements and Galvao (2008) to forecast quarterly GDP. It differs from the MIDAS model proposed by Ghysels (2004) because it does not use Almon-distributed lag functions, and thus does not restrict its parameters. The explanation on Almon-distributed lags will follow later.

The U-MIDAS model simply models the low-frequency variables by including (at least) all the high-frequency explanatory variables corresponding to the low-frequency same period. Throughout this paper we will assume, for simplicity, that there is only one explanatory variable that is included with multiple lags. Franses (2016) denotes the difference between two frequencies as S (i.e. months to quarters indicates $S = 3$, weeks to years indicates $S = 52$). That allows for the definition in (2.1), in which Y_T is a low-frequency independent variable and $X_{s,T}$ is the s^{th} explanatory observation from the low-frequency period T, $s \in S$. In this definition k is the number of lags of the explanatory variable, and κ is some period dependent on the value of k.

$$Y_T = \beta_0 X_{S,T} + \beta_1 X_{S-1,T} + \cdots + \beta_{k-1} X_{S-k,\kappa} \tag{2.1}$$

The U-MIDAS model will be used in replication of work done by Franses (2016) in order to assess the workings and accuracy of the U-MIDAS model. This is a necessary step in order to be justify the use of different kinds of U-MIDAS models.

2.2.1 Yearly to Biannual

In order to properly test the functioning of the U-MIDAS models a simulation experiment will be performed. This simulation uses the DGP specified in (2.2), with $\varepsilon_t \sim N(0, 1)$. This DGP is then treated as if we only observe the sum of two consequent observations half the time, making it a lower frequency variable. As such, a discrepancy between explanatory and dependent variables arises in terms of rate of occurrence. The creation

of the low-frequency variable is done using the *HILO* transformation proposed by Franses (2016): the aggregate low-frequency variables (Y_T) are defined as the sum (average) of the corresponding high frequency flow (stock) variables (y_t, also denoted $Y_{i,T}$ with $i \in S$)[1]. The first step of Franses' *HILO* transformation for biannual observations is given in (2.3).

After consequently multiplying both sides of (2.3) with the inverse of the utmost left matrix and the vector $[1\ 1]$ we arrive at the low frequency definition, given by (2.4). This model, however, still includes an unobserved variable ($Y_{2,T}$). But, as (2.5) holds, the model specified in (2.6) can be used to perform the U-MIDAS regression on the generated Y_T values[2]. (2.5) is the U-MIDAS model proposed by Franses (2016). The true value of the parameters in (2.6) can be found in Table 2.1.

$$y_t = \alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t \tag{2.2}$$

$$\begin{pmatrix} 1 & 0 \\ -\alpha & 1 \end{pmatrix} \begin{pmatrix} Y_{1,T} \\ Y_{2,T} \end{pmatrix} = \begin{pmatrix} 0 & \alpha \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Y_{1,T-1} \\ Y_{2,T-1} \end{pmatrix} + \begin{pmatrix} \beta_0 & 0 \\ \beta_1 & \beta_0 \end{pmatrix} \begin{pmatrix} X_{1,T} \\ X_{2,T} \end{pmatrix} + \begin{pmatrix} 0 & \beta_1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X_{1,T-1} \\ X_{2,T-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,T} \\ \varepsilon_{2,T} \end{pmatrix} \tag{2.3}$$

$$\begin{gathered} Y_T = Y_{2,T} + Y_{1,T} = (\alpha + \alpha^2) Y_{2,T-1} + \\ \beta_0 X_{2,T} + (\beta_0 + \alpha\beta_0 + \beta_1) X_{1,T} + (\beta_1 + \alpha\beta_1) X_{2,T-1} + (\alpha + 1)\varepsilon_{1,T} + \varepsilon_{2,T} \end{gathered} \tag{2.4}$$

$$\alpha Y_{2,T-1} = \alpha(\alpha Y_{1,T-1} + \beta_0 X_{2,T-1} + \beta_1 X_{1,T-1} + \varepsilon_{2,T-1}) \tag{2.5}$$

$$Y_T = \mu + \rho Y_{T-1} + \delta_0 X_{2,T} + \delta_1 X_{1,T} + \delta_2 X_{2,T-1} + \delta_3 X_{1,T-1} + u_t \tag{2.6}$$

Table 2.1: The true values of the parameters in the U-MIDAS regression for biannual to annual data (2.6)

Parameter	True value		
ρ	α^2		
δ_0	β_0		
δ_1	$(1+\alpha)\beta_0$	+	β_1
δ_2	$\alpha\beta_0$	+	$(1+\alpha)\beta_1$
δ_3			$\alpha\beta_1$

[1]Note that flow variables are variables that concern a 'flow', and as such are to be summed in order to aggregate, where a stock variable gives a snapshot of a situation and as such should be averaged in order to aggregate different observations.

[2]The lagged MA term is omitted as Franses (2016) shows it has little added value in the model.

2.2.2 Yearly to Quarterly

In order to further asses the quality of the U-MIDAS model proposed by Franses, a simulation has been performed using the DGP specified in (2.7). This DGP can be used to construct low frequency observations with $S = 4$, which resembles a yearly to quarterly U-MIDAS model. After performing a *HILO* transformation we can define a U-MIDAS regression in a similar way as in Section 2.2.1. The U-MIDAS regression for this simulation is thus defined as in (2.8). The true values of the parameters in (2.8) can be found in Table 2.2.

The assessment that the U-MIDAS definition produces accurate estimations for different levels of data aggregation will validate the search for for the optimal aggregation level. Therefore it is valuable to study the accuracy of U-MIDAS models with different levels of aggregation of data.

$$y_t = \alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \varepsilon_t \tag{2.7}$$

$$\begin{aligned} Y_T =& Y_{4,T} + Y_{3,T} + Y_{2,T} + Y_{1,T} = \mu + \rho Y_{T-1} + \\ & \beta_0^* x_{4,T} + \beta_1^* x_{3,T} + \beta_2^* x_{2,T} + \beta_3^* x_{1,T} + \\ & \beta_4^* x_{4,T-1} + \beta_5^* x_{3,T-1} + \beta_6^* x_{2,T-1} + \beta_7^* x_{1,T-1} + \\ & \beta_8^* x_{4,T-2} + \beta_9^* x_{3,T-2} + \varepsilon_T \end{aligned} \tag{2.8}$$

TABLE 2.2: The true values of the parameters in the U-MIDAS regression for quarterly to annual data, described in (2.8)

Parameter	True value
ρ	α^4
δ_0	β_0
δ_1	$\beta_0(1+\alpha) + \beta_1$
δ_2	$\beta_0(1+\alpha+\alpha^2) + \beta_1(1+\alpha) + \beta_2$
δ_3	$\beta_0(1+\alpha+\alpha^2+\alpha^3) + \beta_1(1+\alpha+\alpha^2) + \beta_2(1+\alpha) + \beta_3$
δ_4	$\beta_0(\alpha+\alpha^2+\alpha^3) + \beta_1(1+\alpha+\alpha^2+\alpha^3) + \beta_2(1+\alpha+\alpha^2) + \beta_3(1+\alpha)$
δ_5	$\beta_0(\alpha^2+\alpha^3) + \beta_1(\alpha+\alpha^2+\alpha^3) + \beta_2(1+\alpha+\alpha^2+\alpha^3) + \beta_3(1+\alpha+\alpha^2)$
δ_6	$\beta_0\alpha^3 + \beta_1(\alpha^2+\alpha^3) + \beta_2(\alpha+\alpha^2+\alpha^3) + \beta_3(1+\alpha+\alpha^2+\alpha^3)$
δ_7	$\beta_1\alpha^3 + \beta_2(\alpha^2+\alpha^3) + \beta_3(\alpha+\alpha^2+\alpha^3)$
δ_8	$\beta_2\alpha^3 + \beta_3(\alpha^2+\alpha^3)$
δ_9	$\beta_3\alpha^3$

2.3 MIDAS Models

2.3.1 Almon Distributed Lag Model

In order to be able to use highly different frequencies of observations, we need to account for the problem of having to estimate a lot of parameters. When using a model to fit weekly data onto yearly observations, and assuming the 'original', high-frequency DGP contains four significant x_t lags, one would theoretically need to estimate 106 parameters. We will denote this number of parameters to be estimated in a U-MIDAS specification as m. As we usually do not have datasets spanning over 100 years, this poses a problem. Ghysels (2004) proposed to use the lag structure proposed by Almon (1965) (Almon lags) in order to account for this problem. As Ghysels was the inventor of the whole MIDAS concept, this model will be denoted as the MIDAS model and the unrestricted model will be referred to as the U-MIDAS model. Currently Almon lags are a widely accepted technique to decrease the required number of parameters to estimate. Almon used Weierstrass's Approximation Theorem, which tells us that:

> "*Every continuous function defined on a closed interval* $[a, b]$ *can be uniformly approximated, arbitrarily closely, by a polynomial function of finite degree,* P*.*"

As such Ghysels suggests to restrict the parameters of the MIDAS model to force them to lie on a polynomial of degree P, as in (2.9). This alleviates our number of parameters to be estimated from k to P. Usually, P is assigned a (rather) small value, such as $2, 3$, or 4, and $P < m$. However, there does not exist a way to properly determine the best degree and lag of the polynomial other than trial and error. Schmidt and Waud (1973) warn their audience about the results of misspecification when using Almon lags and Frost (1975) shows that using the maximization of corrected R-squared values results in biased and non-normal estimators for the parameters. Pagano and Hartley (1981) suggest to use a two-step approach for choosing the correct number of lags and degrees when using the Almon lags, in which first the optimal number of lags is determined using the Akaike Information Criterion (AIC), and then consequently using AIC to determine the optimal degree of the polynomial.

$$\beta_i = \frac{\alpha_0 + \alpha_1 i + \alpha_2 i^2 + \cdots + \alpha_P i^P}{\sum_{j=1}^{I} \alpha_0 + \alpha_1 j + \alpha_2 j^2 + \cdots + \alpha_P j^P} \tag{2.9}$$

2.3.2 Simulation and Data Application

To test whether the optimal level of data aggregation can be determined in advance of making forecasts, both simulation and real-life data will be used to construct forecasts on different levels of data aggregation. Both experiments will consist of weekly data that is aggregated to forecast yearly observations, either with a weekly, monthly, quarterly, biannually or yearly frequency. These models are described in a comparable fashion to those described in subsections 2.2.1 and 2.2.2. The number of lags included in all different MIDAS models can be found in Table 2.3. Based on a 90% estimation sample, the remaining 10% of the sample will be forecasted recursively through one-step-ahead forecasts. In the simulation a DGP with four x_t lags will be used to generate the data. In order to evaluate their accuracy the models' R^2 values and Akaike Information Criteria will be compared to find which teller best predicts forecasting accuracy. Then the models' Root Mean Squared Prediction Error (RMSPE) performances will be compared to see who performed the best with regards to forecasting.

TABLE 2.3: The number of x lags that should theoretically be included in various MIDAS models using different levels of data aggregation. Theoretical number of x lags included are based on an assumed unobserved DGP (as described by Franses, 2016) containing four lags.

Data aggregation level	No. of lags included
Annual	2
Biannual	4
Quarterly	9
Monthly	26
Weekly	106

Chapter 3

Simulations

In this chapter I first describe simulations ran to confirm that the use of U-MIDAS models is appropriate given the used definition. In this simulation, the U-MIDAS regression definition as specified by Franses (2016) is used (see Chapter 2). First the simulation that is performed by Franses is replicated which models annual data using biannual explanatory variables. Consequently the functionality of the U-MIDAS definition is tested on annual data that is explained by quarterly data. Finally forecasts are constructed based on the simulated data and I attempt to find ways to determine the optimal data aggregation level based on the estimation sample using MIDAS models with Almon lags.

3.1 Explanatory Power of U-MIDAS Models

3.1.1 Experiment Description

First a justification of the use of U-MIDAS models for explaining our data is required. In order to assess the accuracy of the U-MIDAS model I have performed simulation runs for different values of the first y_t lag coefficient (α) and number of low-frequency observations (N) and investigate the values of the bias and standard deviation of the parameter estimators of the regression. If the bias and standard deviation of the regression have acceptably low values one can conclude that it is appropriate to use a U-MIDAS regression to explain annual data using biannual explanatory variables. Then, we might expect that we can also use U-MIDAS models properly to produce forecasts of our low-frequency data.

3.1.2 Biannual Data to Annual Observations

The results of the simulation using simulated biannual data to fit annual observations, introduced in (2.2) to (2.6), can be found in Table 3.1. It is remarkable that the bias of estimations does not necessarily appear to improve with an increase in the number of observations (N). The standard deviation intuitively does decrease with an increase in N. Overall the model appears to estimate the parameters very well and even with small values for α and N the model provides reasonable estimates for most parameters.

TABLE 3.1: Simulation results based on a sample of N 'yearly' observations, when a 'biannual' DGP is the true process with $x_t \sim N(1,1)$, $\varepsilon_t \sim N(0,1)$, and $y_0 \sim N(0,1)$ and $y_t = \alpha y_{t-1} + x_t + 2x_{t-1} + \varepsilon_t$, 1.000 replications. The U-MIDAS regression is: $Y_T = \mu + \rho Y_{T-1} + \delta_0 X_{2,T} + \delta_1 X_{1,T} + \delta_2 X_{2,T-1} + \delta_3 X_{1,T-1} + u_t$

		ρ		δ_0		δ_1		δ_2		δ_3	
α	N	mean	std	mean	std	mean	std	mean	std	mean	std
True		0.25		1		3.5		3.5		1	
0.5	40	0.30	0.14	1.02	1.29	3.47	1.29	3.46	1.30	0.75	1.38
	400	0.34	0.04	1.02	0.38	3.48	0.38	3.41	0.38	0.65	0.40
True		0.64		1		3.8		4.4		1.6	
0.8	40	0.64	0.11	0.97	1.53	3.86	1.53	4.36	1.53	1.68	1.58
	400	0.69	0.03	0.99	0.44	3.80	0.44	4.38	0.45	1.42	0.46
True		0.9025		1		3.95		4.85		1.9	
0.95	40	0.88	0.05	0.99	1.66	3.92	1.65	4.83	1.66	1.94	1.65
	400	0.91	0.01	1.02	0.48	3.97	0.48	4.83	0.48	1.87	0.48

3.1.3 Quarterly Data to Annual Observations

When looking at the results for the quarterly-to-yearly observations in Table 3.2[1] it appears that again different values of N do not improve the parameter bias in any way. The standard deviation however is still, as is intuitive, decreased with an increase in the number of data points. It is very apparent that the parameters for x lags that are directly included in the DGP $(x_{4,T}, \cdots, x_{2,T-1})$ are estimated with smaller biases than those that only influence Y_T observations indirectly through the U-MIDAS structure $(x_{1,T-1}, \cdots, x_{3,T-2})$. Again all observations are acceptably close which defends the use of

[1]Full version of Table 3.2 is available in Appendix A.1.

U-MIDAS models at different frequencies and thus the comparison of the different models could be interesting.

TABLE 3.2: Selection of simulation results based on a sample of N 'yearly' observations, when a 'quarterly' DGP is the true process with $x_t \sim N(1,1)$, $\varepsilon_t \sim N(0,1)$, $y_0 \sim N(0,1)$, DGP is $y_t = \alpha + x_t + 1.2x_{t-1} + 0.8x_{t-2} + 0.7x_{t-3} + \varepsilon_t$, 1.000 replications. The U-MIDAS regression is: $Y_T = \mu + \rho Y_{T-1} + \delta_0 x_{4,T} + \delta_1 x_{3,T} + \delta_2 x_{2,T} + \delta_3 x_{1,T} + \delta_4 x_{4,T-1} + \delta_5 x_{3,T-1} + \delta_6 x_{2,T-1} + \delta_7 x_{1,T-1} + \delta_8 x_{4,T-2} + \delta_9 x_{3,T-2} + \varepsilon_T$
The full version of this Table can be found in Appendix A.1

		$\alpha = 0.5$			$\alpha = 0.8$			$\alpha = 0.95$		
		True	N = 40	N = 400	True	N = 40	N = 400	True	N = 40	N = 400
ρ	mean	0.0625	0.12	0.13	0.41	0.42	0.43	0.81	0.81	0.82
	std		0.13	0.04		0.06	0.02		0.02	0.01
δ_0	mean	1.00	1.00	1.00	1.00	1.07	1.01	1.00	1.03	1.01
	std		0.62	0.17		0.94	0.25		1.18	0.31
δ_1	mean	2.70	2.74	2.69	3.00	2.96	3.01	3.15	3.17	3.14
	std		0.62	4.35		0.94	0.25		1.18	0.31
δ_2	mean	4.35	4.34	4.35	5.40	5.41	5.40	5.99	6.01	5.99
	std		0.62	0.17		0.94	0.25		1.18	0.31
⋮						⋮				⋮
δ_8	mean	0.3625	0.067	-0.037	1.22	1.11	0.99	1.92	2.01	1.87
	std		0.950	0.258		1.08	0.30		1.19	0.32
δ_9	mean	0.0875	-0.130	-0.225	0.36	0.27	0.12	0.60	0.58	0.55
	std		0.832	0.225		1.07	0.29		1.19	0.32

3.2 Forecasting Power of U-MIDAS Models

The results of the forecasts of different models are depicted in Tables 3.3 and 3.4. The accuracy of the models is expressed in terms of Root Mean Squared Prediction Errors (RMSPEs), which contain information on both the bias and the standard deviation of the forecasts. It is immediately obvious that the U-MIDAS model always performs worse than the true DGP, which makes sense. But, the smaller the value of α (the first y_t lag

coefficient) the better the performance of U-MIDAS. This makes sense as the higher the value of α, the larger the advantage of the True DGP of including the y_{t-1} lags. When the discrepancy between frequencies is small (biannual to yearly), the significance of the lagged y_t is small ($\alpha = 0.5$) and there is a realistic sample size ($N = 40$), U-MIDAS' performance even approaches the true DGP in terms of RMSPE.

TABLE 3.3: RMSPE values for different values of α and N, using the biannual DGP specified in (2.2). Hold-out sample consists of $\lceil 10\% \rceil$ of the observations.

	$N = 40$		$N = 400$	
	U-MIDAS	True DGP	U-MIDAS	True DGP
$\alpha = 0.5$	6.555	6.163	6.503	4.167
$\alpha = 0.8$	9.098	4.044	10.049	4.430
$\alpha = 0.95$	15.579	4.431	9.789	4.504

TABLE 3.4: RMSPE values for different values of α and N, using the quarterly DGP specified in (2.7). Hold-out sample consists of $\lceil 10\% \rceil$ of the observations.

	$N = 40$		$N = 400$	
	U-MIDAS	True DGP	U-MIDAS	True DGP
$\alpha = 0.5$	2.899	1.401	3.193	1.594
$\alpha = 0.8$	5.128	1.333	5.486	1.703
$\alpha = 0.95$	7.928	1.290	6.404	1.394

3.3 Optimal Level of Data Piling

In order to assess whether there is an added value to looking at different levels of data aggregation, a simulation has been performed in which the high-frequency y_t is sampled at a weekly frequency. Through a *HILO* transformation, the low-frequency Y_T observations are generated that are then regressed on different levels of aggregated explanatory

variables, x. A comparison is made in terms of RMSPE performance and as predictors for RMSPE performance R^2 and AIC are considered.

First off, the appropriate Almon degree for this simulation had to be selected. Table 3.5 displays how many times per level of data aggregation a certain degree was optimal (degrees were considered between 2 and 6). It is apparent that an Almon degree of 2 most often is the best fit for this model. Therefore all other values in this section have been generated using MIDAS models with Almon lags of degree 2.

Table 3.6 is a cross table comparing the occurrence of optimality in terms of RMSPE to that of optimality in R^2 for different levels of data aggregation. It immediately stands out that the model using annual observations of the explanatory variables never produces the optimal forecasts. This is probably due to the large loss in information that is incurred by aggregating all (independent) explanatory variables. It is also remarkable that the diagonal entries are all the largest in their respective rows. This shows that R^2 optimality coincides with forecasting optimality more often (on average 37% of the cases) than that it occurs with any other frequency's optimality.

Table 3.7 is a cross table comparing the occurence of optimality in terms of RMSPE to that of optimality in AIC for different levels of data aggregation. The Table shows similar results to those of Table 3.6 except the diagonal entry for the biannual level of data aggregation is not the highest in its row. It appears to be the case that the AIC overvalues the inclusion of (much) information.

TABLE 3.5: Cross table showing the occurrence of (forecasting performance) optimality for different frequencies and different Almon degrees. Results are based on a simulation experiment with 1,000 replications. The DGP used is as in (2.7) with $\varepsilon \sim N(0,1)$.
DGP: $y_t = 0.8y_{t-1} + x_t + 1.2x_{t-1} + 0.8x_{t-2} + 0.7x_{t-3} + \varepsilon_t$

		Almon degree				
		2	3	4	5	6
Frequency	Annual	487	261	163	67	22
	Biannual	653	346	1	0	0
	Quarterly	431	165	133	124	147
	Monthly	407	166	142	135	150
	Weekly	396	181	152	135	136
		2,374	1,119	591	461	455

TABLE 3.6: Cross table showing the occurrence of (forecasting performance) optimality for different frequencies versus R^2 optimality for different frequencies. Results are based on a simulation experiment with 1,000 replications. The DGP used is as in (2.7) with $\varepsilon \sim N(0,1)$.
DGP: $y_t = 0.8y_{t-1} + x_t + 1.2x_{t-1} + 0.8x_{t-2} + 0.7x_{t-3} + \varepsilon_t$

		R^2 *Optimality*				
		Annual	Biannual	Quarterly	Monthly	Weekly
Forecasting Optimality	Annual	0	0	0	0	0
	Biannual	2	102	66	71	75
	Quarterly	7	47	88	46	63
	Monthly	4	49	40	70	52
	Weekly	3	48	39	35	93
		16	246	233	222	283

TABLE 3.7: Cross table showing the occurrence of (forecasting performance) optimality for different frequencies versus AIC optimality for different frequencies. Results are based on a simulation experiment with 1,000 replications. The DGP used is as in (2.7) with $\varepsilon \sim N(0,1)$.
DGP: $y_t = 0.8y_{t-1} + x_t + 1.2x_{t-1} + 0.8x_{t-2} + 0.7x_{t-3} + \varepsilon_t$

		AIC Optimality				
		Annual	Biannual	Quarterly	Monthly	Weekly
Forecasting Optimality	Annual	0	0	0	0	0
	Biannual	0	40	91	88	97
	Quarterly	0	19	104	54	74
	Monthly	0	15	53	81	66
	Weekly	0	19	49	40	110
		0	93	297	263	347

Chapter 4

Application on Unemployment and Staffing Data

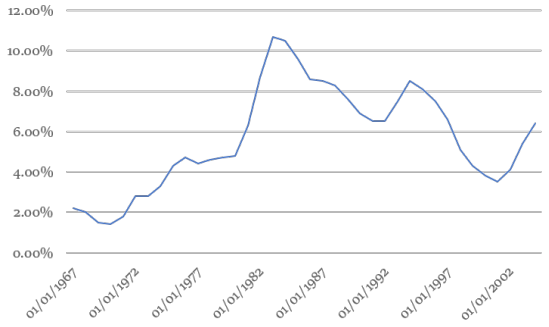
This section describes the application of the aforementioned techniques on a real-life dataset. By doig so I am able to assess whether the forecasting using different levels of aggregated data also produces different results in empirical environments.

4.1 Data Description

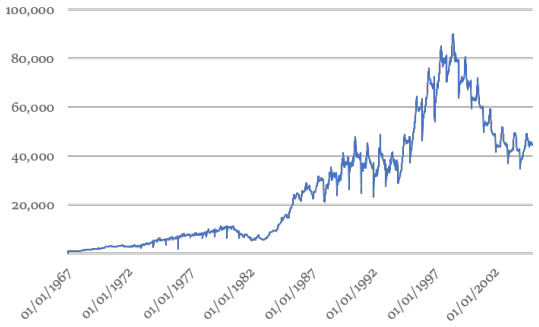
The dependent data used is this research concerns the levels of unemployment in the Netherlands originating from the Centraal Bureau voor Statistiek (CBS, Dutch Central Bureau for Statistics) and independent data about the temporary workers under contract with Randstad, the biggest Dutch staffing agency. The data was gathered between 1967 and 2004, since in 2005 the definition for Randstad's data was altered, which caused a break in the data. Furthermore, the first year of observations was lost in the data transformation process, which leaves 37 yearly observations about unemployment and 1.924 weekly observations about Randstad employees. For visualization purposes, the raw data regarding Randstad and unemployment are displayed in Figure 4.1.

4.1.1 Data Transformations

Obviously, different levels of data piling had to be constructed. For the aggregation from weeks to months, $4\frac{1}{3}$ weeks had to be put into each month, which means some weeks

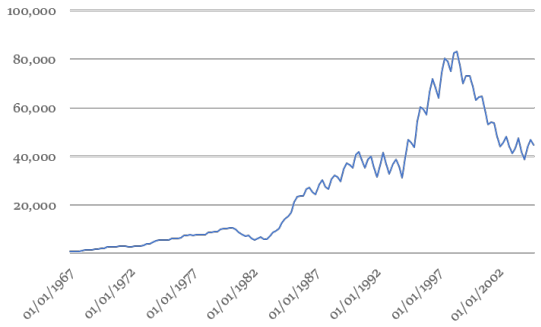


(A) Yearly observations on the unemployment rate in the Netherlands

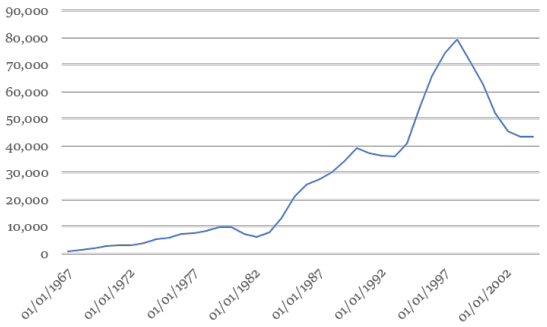


(B) Weekly observations on the number of people under contract with Randstad B.V.

FIGURE 4.1: Raw data inputs for the application section



(A) Quarterly observations on the number of people under contract with Randstad B.V.



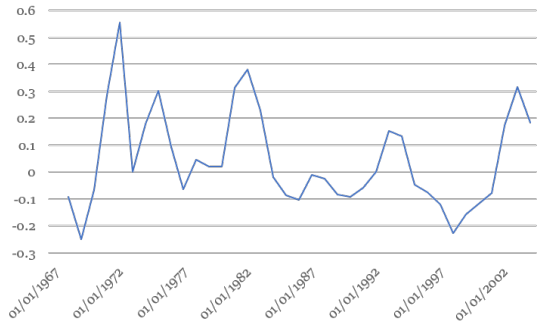
(B) Yearly observations on the number of people under contract with Randstad B.V.

FIGURE 4.2: Different frequencies of the number of people under contract with Randstad constructed through data piling.

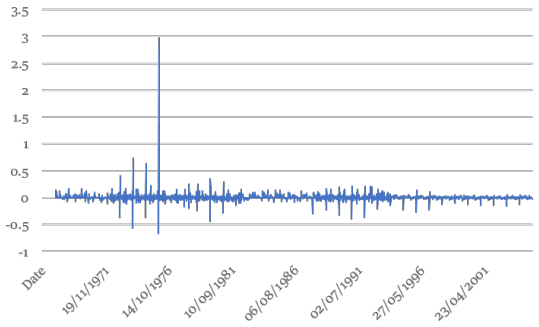
were put into one month for $\frac{1}{3}^{th}$ and in the other month for $\frac{2}{3}^{th}$. Some results of data piling can be found in Figures 4.2a and 4.2b. Furthermore, in order to account for the non-stationarity embedded in the Randstad data, the data has been transformed to first differences rather than absolute values, see Figure 4.3.

4.2 Application Results

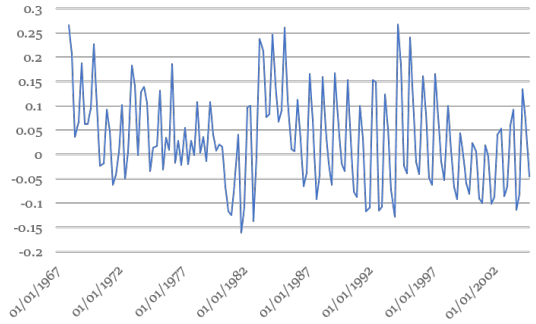
Figure 4.4 shows the R^2 values, AIC values, and RMSPEs for U-MIDAS models explaining the annual unemployment data in the Netherlands using quarterly, biannual, and annual aggregations of the weekly available number of Randstad payroll jobs. The optimal values for each of the series are darkly outlined in all figures for graphical purposes. There are no restrictions imposed on the shape of the parameter curve. It is remarkable that the biannual data aggregation level outperforms both other models in terms of R^2, AIC and



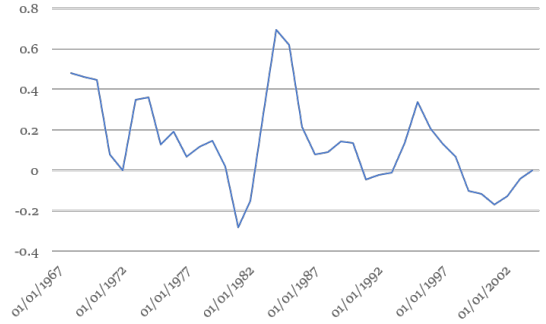
(A) First differences of Dutch unemployment data



(B) First difference of the number of people under contract with Randstad B.V. at a weekly frequency



(C) First difference of the number of people under contract with Randstad B.V. at a quarterly frequency



(D) First difference of the number of people under contract with Randstad B.V. at an annual frequency

FIGURE 4.3: First differences of the graphs in Figures (4.2) and (4.1)

RMSPE. This result is coherent with the hypothesis that there might be a correlation between the models' in-sample explanatory power and out-of-sample forecasting power. Due to the limited number of years of data and the high number of parameters to be estimated, though, it is not possible to apply weekly or monthly data aggregation levels on this method. As such the model parameters are restricted by using an Almon Distributed Lag (Almon) model.

However, Figures 4.5a and 4.5b show that for some of the most common Almon degrees ($P = 2, P = 4$) the best R^2 and AIC values do not coincide with the best RMSPE values. Also the AIC values have been scaled. For $P = 2$ the best RMSPE even coincides with the worst R^2 value. These results appear to contradict our expected outcomes, however, they can arise due to misspecification of the MIDAS model with Almon lags (Schmidt & Waud, 1973). Actually, when looking at the AIC selection criterion specified by Pagano and Hartley (1981), see also Figure 4.6, it becomes clear that the optimal Almon degree for this dataset is $P = 3$.

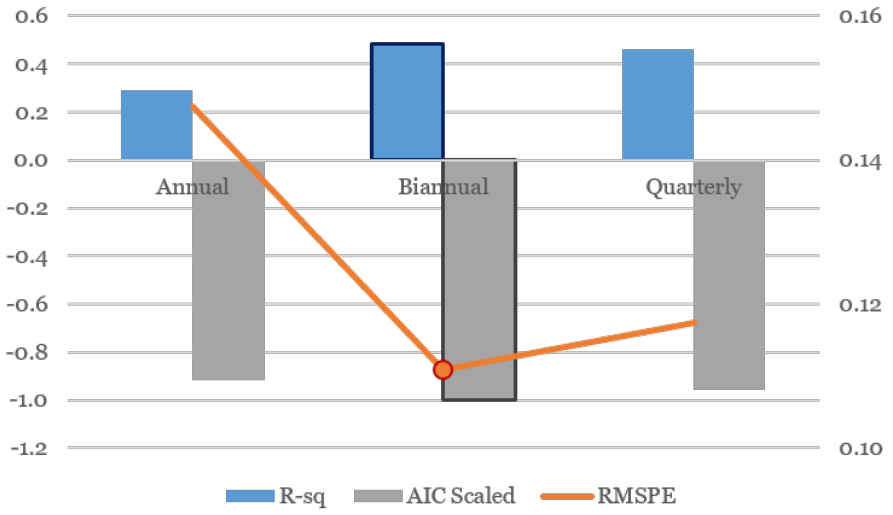
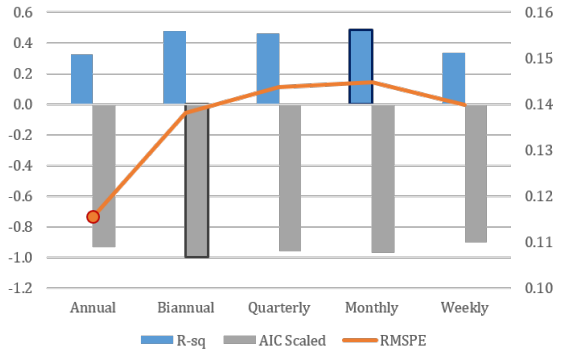
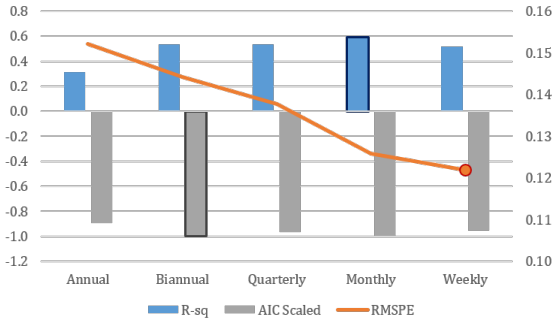


FIGURE 4.4: Results for the Dutch unemployment data forecasted using a MIDAS model with beta polynomial. Given in the figure are the R^2 values (LHS) of the estimation sample and the RMSPEs (RHS) of the forecasting sample for annual, biannual, and quarterly observations.



(A) MIDAS Model with Almon Distributed Lag with $P = 2$



(B) MIDAS Model with Almon Distributed Lag with $P = 4$

FIGURE 4.5: Results for the Dutch unemployment data forecasted using a MIDAS model with Almon Distributed Lag polynomials. Given in the figure are the R^2 values and scaled versions of the AIC (LHS) of the estimation sample and the RMSPEs (RHS) of the forecasting sample for annual, biannual, quarterly, monthly, and weekly observations. The estimation sample consists of 33 observations.

In the optimal model according to the AIC selection criterion (with $P = 3$) the lowest RMSPE occurs with a biannual level of data aggregation. Furthermore, the R^2 is at its maximum and the AIC at its minimum for the biannual level of data aggregation. In this case, model selection based on either R^2 or AIC would thus have resulted in improved forecasting power of the model. In comparison, the forecasts made by the biannual model with $P = 3$ are 10% and 15% more accurate than weekly and annual models, respectively.

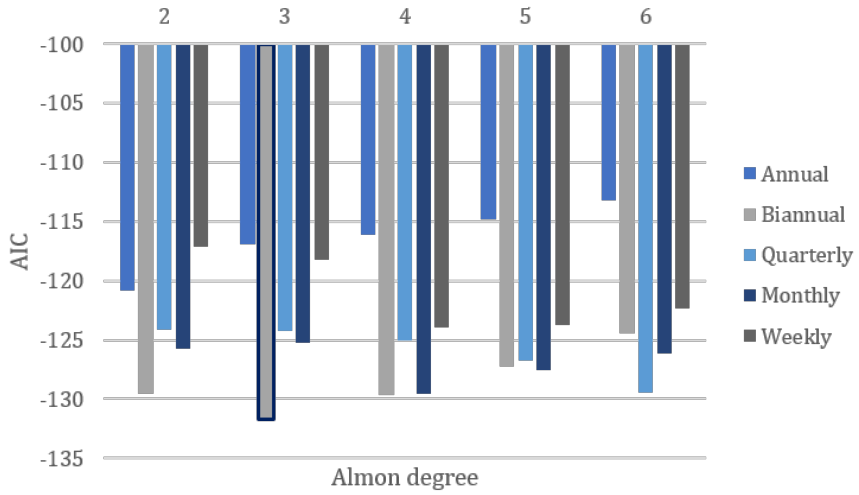


FIGURE 4.6: Results for the Dutch unemployment data forecasted using a MIDAS model with different degrees of Almon lags. Given in the figure are the AIC values of the estimation sample annual, biannual, and quarterly observations.

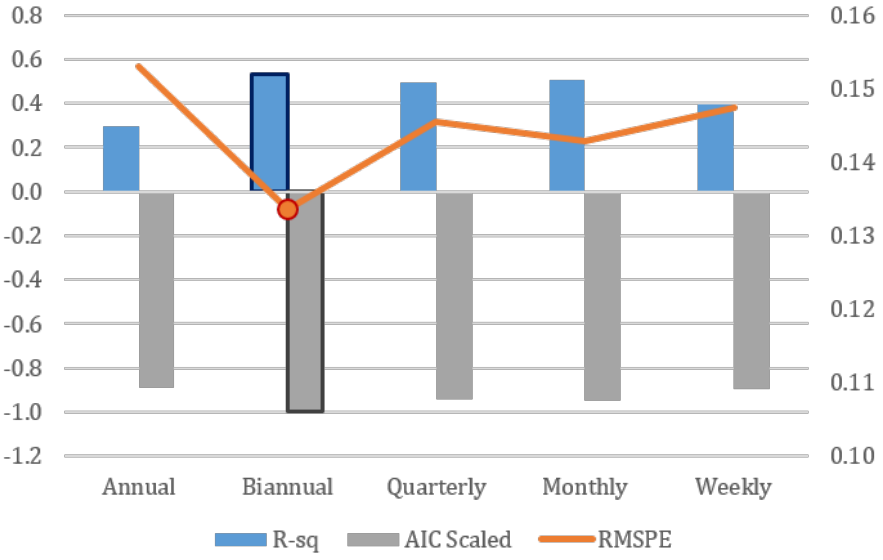


FIGURE 4.7: Results for the Dutch unemployment data forecasted using a MIDAS model with Almon Distributed Lag polynomial of degree 3. Given in the figure are the R^2 values and scaled versions of the AIC (LHS) of the estimation sample and the RMSPEs (RHS) of the forecasting sample for annual, biannual, quarterly, monthly, and weekly observations.

Chapter 5

Conclusion

This research is focused on studying the possible added value of data piling in the use of MIDAS models. Results of models using intermediate levels of data aggregation (monthly, quarterly and biannual aggregation in the Application chapter) are compared to the 'traditional' models that are currently used most commonly (annual and weekly aggregation in the Application chapter). Simulation results show that more often than not the aggregation of data can lead to an improvement in both in-sample explanatory power and out-of-sample forecasting power. Furthermore, in some cases it appears that R^2 and AIC optimality in the in-sample regression are valid indicators of forecasting optimality in the hold-out sample. However, there remain a notable amount of cases in which these tellers do not correctly predict which level of data aggregation is optimal. As such, it would prove valuable to further research the possible indicators of forecasting optimality with regards to data aggregation level selection.

What can be concluded is that oftentimes data aggregation does improve explanatory and forecasting power and, as such, there lies value in the evaluation of differences between forecasts of different models (i.e. different levels of data aggregation). As is demonstrated in the Application chapter, the aggregation of data can lead to worthy improvements of the forecasting power of the MIDAS model, whose results show 15% more accurate forecasts than the traditional fully aggregated model (annual-to-annual modeling) and also 10% more accurate forecasts than the well-known MIDAS model that does not aggregate data (weekly-to-annual modeling). In this particular example the optimality of forecasting accuracy, R^2 and AIC happen to coincide.

In conclusion, the data piling process appears to be a valuable addition to the MIDAS method for modeling time series but the challenge will remain to find a way to properly

predict the optimal data aggregation level. Further research can prove valuable if a proper technique were to be found to make this prediction.

Appendix A

Tables

TABLE A.1: Simulation results based on a sample of N 'yearly' observations, when a 'quarterly' DGP is the true process with $x_t \sim N(1,1)$, $\varepsilon_t \sim N(0,1)$, $y_0 \sim N(0,1)$, DGP is $y_t = \alpha + x_t + 1.2x_{t-1} + 0.8x_{t-2} + 0.7x_{t-3} + \varepsilon_t$ (1000 replications)

		$\alpha = 0.5$			$\alpha = 0.8$			$\alpha = 0.95$		
		True	N = 40	N = 400	True	N = 40	N = 400	True	N = 40	N = 400
γ	mean	0.0625	0.12	0.13	0.41	0.42	0.43	0.81	0.81	0.82
	std		0.13	0.04		0.06	0.02		0.02	0.01
δ_0	mean	1.00	1.00	1.00	1.00	1.07	1.01	1.00	1.03	1.01
	std		0.62	0.17		0.94	0.25		1.18	0.31
δ_1	mean	2.70	2.74	2.69	3.00	2.96	3.01	3.15	3.17	3.14
	std		0.62	4.35		0.94	0.25		1.18	0.31
δ_2	mean	4.35	4.34	4.35	5.40	5.41	5.40	5.99	6.01	5.99
	std		0.62	0.17		0.94	0.25		1.18	0.31
δ_3	mean	5.875	5.86	5.87	8.02	8.03	8.02	9.39	9.43	9.38
	std		0.63	0.17		0.94	0.25		1.19	0.31
δ_4	mean	5.58	5.53	5.50	8.71	8.79	8.69	10.81	10.81	10.80
	std		0.63	0.17		0.94	0.25		1.18	0.31
δ_5	mean	4.53	3.98	3.97	7.85	7.49	7.50	10.07	10.01	9.96
	std		0.72	0.19		0.95	0.26		1.18	0.31
δ_6	mean	2.59	2.39	2.28	5.52	5.50	5.38	7.73	7.74	7.70
	std		0.83	0.23		0.99	0.27		1.17	0.31
δ_7	mean	1.06	0.74	0.64	2.90	2.87	2.70	4.33	4.38	4.30
	std		0.979	0.265		1.04	0.29		1.19	0.32
δ_8	mean	0.3625	0.067	-0.037	1.22	1.11	0.99	1.92	2.01	1.87
	std		0.950	0.258		1.08	0.30		1.19	0.32
δ_9	mean	0.0875	-0.130	-0.225	0.36	0.27	0.12	0.60	0.58	0.55
	std		0.832	0.225		1.07	0.29		1.19	0.32

Bibliography

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1):178–176.

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

Clements, M. and ao, A. G. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the united states. *Journal of Business & Economic Statistics*, 26(4):546–554.

Foroni, C., Marcellino, M., and Schumacher, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):57–82.

Franses, P. H. B. F. (2016). Yet another look at MIDAS regression. (No. EI2016-32). *Econometric Institute Research Papers*.

Frost, P. (1975). Some properties of the almon lag technique when one searches for degree of polynomial and lag. *Journal of the American Statistical Association*, 70(351):606–612.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. CIRANO Working paper 2004s-20.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1):59–95.

Koenig, E., Dolmas, D., and Piger, J. (2003). The use and abuse of real-time data in economic forecasting. *Review of Economics and Statistics*, 85(3):618–628.

Monteforte, L. and Moretti, G. (2013). Real-time forecasts of inflation: The role of financial variables. *Journal of Forecasting*, 32(1):51–61.

Pagano, M. and Hartley, M. J. (1981). On fitting distributed lag models subject to polynomial restrictions. *Journal of the Econometrics*, 2(16):171–198.

Schmidt, P. and Waud, R. N. (1973). The almon lag technique and the monetary versus policy debate. *Journal of the American Statistical Association*, 68(341):11–19.