Modeling non-stationary and stationary mixed-frequency time series

Nederlandse samenvatting

De frequentie waarmee economische tijdreeksen worden waargenomen verschilt per variabele. Zo wordt het bruto binnenlands product ieder kwartaal gepubliceerd, terwijl veel economische indicatoren iedere maand of nog vaker verschijnen. Deze variatie in verschijningsfrequentie zorgt voor problemen voor econometristen, die moeten besluiten hoe gebruik te maken van de variabelen die vaker waargenomen worden. Een mogelijkheid is om alle data met hoge frequentie direct te gebruiken zonder beperkingen, waarbij de verschillen in frequentie tussen de data effectief genegeerd worden. Deze strategie wordt echter onaantrekkelijk als de verschillen in frequentie groot worden or als er veel variablelen zijn, omdat in dit geval erg veel parameters geschat moeten worden terwijl men daarvoor slechts een beperkt aantal waarnemingen ter beschikking heeft. Als alternatief kan men de data met hoge frequentie aggregeren, ofwel samenvoegen, danwel de data met lage frequentie splitsen, om tot een gemeenschappelijke frequentie voor alle variabelen te komen. Binnen de eerste aanpak kan onderscheid gemaakt worden tussen een vaststaande aggregatie-methode of een die van de data afhangt, waarbij deze tweede optie bekend is geworden als *MI(xed) DA(ta) S(ampling)*.

Ondanks dat de literatuur over dit onderwerp sinds de introductie van MIDAS modellen snel gegroeid is, zijn er maar weinig artikelen die er niet van uitgaan dat de betreffende variabelen *stationair* zijn, wat wil zeggen dat de eigenschappen van de tijdreeksen niet veranderen over de tijd. Aangezien het van veel (macro-)economische variabelen, waargenomen met verschillende frequenties, bekend is dat ze niet stationair zijn, is dit een dubieze aanname. Tijdreeksen zoals bruto binnenlands product, inflatie, nationaal inkomen en wisselkoersen zijn niet stationair en kunnen enkel getransformeerd worden tot stationair door het verschil

tussen hun huidige waarde en de waarde in de vorige periode te berekenen. Ook zijn er veel tijdreeksen die dezelfde veranderlijke eigenschappen delen zoals een gezamenlijke trend, wat *coïntegratie* genoemd wordt en betekent dat er een bepaalde (lange termijn) relatie is tussen deze variabelen die wel stationair is.

Naast de bovengenoemde focus op stationaire tijdreeksen was de literatuur over modellen met variërende frequenties voornamelijk beperkt tot enkelvoudige regressie-modellen. Recentelijk is echter meer aandacht gekomen voor het ontwikkelen van systeem-modellen, in het bijzonder *vector autoregressive (VAR)* modellen, voor processen van variërende frequentie. Echter, net als in de enkelvoudige modellen worden de variabelen nog steeds verondersteld stationair te zijn. Daarnaast zijn veel van de problemen die in meervoudige modellen voorkomen nog niet onderzocht in de context van modellen met variërende frequenties, zoals bijvoorbeeld de causale verbanden tussen variabelen wiens frequentie flink verschilt.

Dit proefschrift complementeert de bestaande literatuur over tijdreeksen met variërende frequenties op de manier zoals beschreven in de vorige twee alinea's. In het bijzonder behandelt dit proefschrift het voorspellen van mogelijk niet-stationaire en gecoïntegreerde tijdreeksen gebruik makende van modellen met variërende frequentie in zowel een enkelvoudig regressie- als een systeem-kader. Wat betreft stationariteit van de variabelen kan onderscheid gemaakt worden tussen hoofdstukken 2 tot 4 die niet-stationaire en mogelijk gecoïntegreerde gevallen bekijken en hoofdstukken 5 en 6 die enkel stationaire reeksen toelaten. Wat betreft het enkelvoudige model ten opzichte van het systeem model, kan onderscheid gemaakt worden tussen hoofdstukken 2 en 3 die het eerste geval en hoofdstukken 4 tot 6 die het tweede geval beschouwen.

Hoofdstuk 2 introduceert een *error-correction* model (ECM) voor niet-stationaire tijdreeksen met variërende frequenties, waar speciale aandacht uitgaat naar de vorm van de lange termijn relatie en diens invloed op de overige, korte termijn, variabelen. In het bijzonder wordt aangetoond dat door een MIDAS model te gebruiken de onderzoeker zich gemakkelijk kan weren tegen dynamische misspecificatie, terwijl de informatie die in het proces met hoge frequentie besloten ligt bewaard blijft.

Deze ECM-MIDAS modellen worden gebruikt in hoofdstuk 3, waarin de analyse van tijdreeksen met variërende frequenties en van *real-time* data worden gecombineerd. Specifiek wordt de *repeated observation forecasting* methode van Stark and Croushore (2002) uitgebreid om verklarende variabelen mee te nemen, die mogelijk met een hogere frequentie worden waargenomen dan de afhankelijke variabele.

Bovendien wordt geïllustreerd door middel van een specifiek schema van veranderlijke gewichten, hoe informatie van het revisie-proces kan worden meegenomen in real-time toepassingen.

Hoofdstuk 4 kan worden gezien als de multivariate tegenhanger van hoofdstuk 2 door het introduceren van een *vector error-correction models (VECM)* met variërende frequenties, waarmee het VAR model van Ghysels (2012) wordt uitgebreid naar niet-stationaire tijdreeksen. Er wordt aangetoond dat een onderzoeker rekening moet houden met twee groepen van lange termijn relaties, en dat zelfs in afwezigheid van coïntegratie tussen de variabelen met lage en hoge frequentie het VAR model met variërende frequenties kan leiden tot verliezen in efficiëntie. Gebruik makend van de VECM representatie met variërende frequenties breiden we de analyse van gemeenschappelijke cyclische patronen van, bijvoorbeeld, Engle and Kozicki (1993) uit naar de situatie met variërende frequenties. Vergeleken met temporeel geaggregeerde modellen, hebben modellen met variërende frequenties niet alleen voordelen wat betreft het doen van voorspellingen, maar ook wat betreft de detectie van gemeenschappelijke cycli.

Hoofdstukken 5 en 6 bekijken het toetsen op causaliteit in een VAR model met variërende frequenties. Terwijl *nowcasting* causaliteit en diens impact op het toetsen op causaliteit in enkelvoudige regressie-modellen met variërende frequentie wordt besproken in het eerste hoofdstuk, is het toetsen op Granger causaliteit in grote VAR modellen met variërende frequenties onderwerp van het laatste hoofdstuk. Meer specifiek, doordat er grote verschillen zijn in frequenties van de variabelen, worden er twee methodes bekeken om het aantal parameters dat geschat moet worden te verminderen, te weten beperkingen op de gereduceerde rang en Bayesiaanse technieken. Voor deze laatste aanpak wordt getoond hoe a priori opvattingen kunnen worden opgenomen in een a priori distributie in de setup met variërende frequenties, waarmee de weg wordt vrijgemaakt voor diepere analyse van Bayesiaanse VAR modellen met variërende frequenties in de toekomst.

Summary Ph.D. thesis by Thomas B. Götz

Time series data are available at mixed frequencies. This simple fact did not only serve as a foundation for an entire branch of econom(etr)ic research, but also for this thesis. In view of the huge amount of distinct data that is available to researchers nowadays, one may alter the initial sentence to 'time series data are available at very mixed frequencies'. Indeed, CO2 emissions,[1] health outlooks or national income accounts appear on an annual basis, national account variables, e.g., the gross domestic product, are often available at a quarterly frequency, many economic indicators, such as the industrial production index or the unemployment rate, get supplied each month, money stock measures provide an example of weekly data, whereas many financial data get published daily, intra-daily or even on a tick-by-tick basis.

When attempting to perform an econometric analysis of variables that are sampled at mixed frequencies, one faces the dilemma of deciding how to make use of the high-frequency observations at one's disposal. On the one extreme, one may choose to simply use all of these observations directly. While this approach may be feasible for small frequency discrepancies among the variables under consideration, it is less attractive if the variables become available in time intervals of greater difference. Not only do the high-frequency observations become more 'noisy' as a variable's frequency increases, researchers quickly face the issue of parameter proliferation, i.e., the amount of parameters becomes too large given usually available sample sizes.

On the other extreme, one can temporally aggregate the high-frequency variables in order to achieve a common low frequency for all time series under consideration. The study of temporal aggregation in the econometric literature dates back almost half a century, with contributions by Amemiya and Wu (1972), Tiao (1972), Nijman and Palm (1990) and Marcellino (1999) among many others. Silvestrini and Veredas (2008) provide an excellent survey on this topic. Due to its simplicity, temporal aggregation has also received a lot of attention in the mixed-frequency literature and is still often considered as a benchmark approach. The basis for the method's simplicity is, however, also its weakness: The temporal aggregation scheme is a priori fixed.[2] If this aggregation scheme differs from the unknown aggregation scheme that is truly underlying the data, researchers may not only discard potentially useful information, they may also end up with biased estimates.

As a means to address this dilemma, Ghysels et al. (2004) introduced MI(xed) DA(ta) S(ampling) models. Their main virtue is their parsimonious specification, which allows researchers to make use of potentially all high-frequency observations with a relatively small number of parameters to estimate. While initial work on MIDAS models has focused on volatility (see Ghysels et al., 2006 or Forsberg and Ghysels, 2007 among others), much work has been devoted to improving quarterly (or monthly) macroeconomic forecasting.[3] Additionally, MI-

[1] At least for data from the World Bank.

[2] Often a simple average of the high-frequency observations or a specific high-frequency observation is chosen as its temporally aggregated counterpart.

[3] Clements and Galvão (2008), Andreou et al. (2010) and Foroni et al. (2012) to only name a few.

DAS has been employed in the context of other regression applications[4] or within dynamic correlation models (Colacito et al., 2011). In this thesis, however, only the former two applications are considered. As far as different extensions and variants of MIDAS regression models are concerned, i.e., different weight polynomial specifications, autoregressive augmentations, non-linear and multivariate MIDAS regression models among others, Ghysels et al. (2007b) provide an excellent overview.

It is worth mentioning that the traditional approach to deal with mixed-frequency data (apart from temporally aggregating the high-frequency variables) has been to interpolate the low-frequency variable to achieve a common high frequency for the variables under consideration. This method rests on the assumption that all data are truly sampled at the high frequency, whereby the high-frequency observations of the variables sampled at lower frequencies are simply missing. The model is then cast in state-space form and the Kalman filter is used to obtain parameter estimates. Handling mixed-frequency data along these lines has been considered in Zadrozny (1988), Mariano and Murasawa (2003) and Giannone et al. (2008) among others.[5] Note that there is a close connection between MIDAS regressions and the Kalman filter in the sense that the former are either exact representations or very accurate approximations of the latter (Bai et al., 2013). In this thesis, though, only MIDAS regressions models (and their system variants) are dealt with.

As far as the econometric analysis of MIDAS regression models is concerned, using a mixed-frequency data generating process, Andreou et al. (2010) derive the asymptotic properties of the nonlinear least squares (NLS) estimator and compare them to the estimator based on a temporally aggregated high-frequency variable. Thereby, the authors show that the latter may indeed result in inefficient, biased and inconsistent estimates, whereby the former is asymptotically more efficient. Further work on the econometric analysis of MIDAS models includes Ghysels et al. (2004), the aforementioned paper by Bai et al. (2013), Kvedaras and Rackauskas (2010), Rodriguez and Puggioni (2010), Wohlrabe (2009) and Foroni (2012), whereby the latter includes a survey of econometric methods for mixed-frequency data.

In all of the aforementioned papers, though, the variables under consideration are either assumed or transformed to satisfy covariance-stationarity, i.e., the first and second moments of the parameters have time-invariant (sample) distributions. In view of the aforementioned frequent application of MIDAS models to macroeconomic forecasting, where many of the associated variables are non-stationary, this may be problematic. Merely transforming all eventually non-stationary variables by taking first (or higher-order) differences may have a fatal effect on the model's forecasting performance due to potentially neglecting a common (long-run) stochastic trend in the variables (Clements and Hendry, 1998). Consequently, an extension of MIDAS regression models to the presence of non-stationarities in the data is called for.[6]

[4] For example, Ghysels et al. (2007a) investigate commercial real estate valuation

[5] Kuzin et al. (2009) compare this approach to MIDAS models in terms of their abilities to now- and forecast GDP growth in the euro area, finding that the two are complements rather than substitutes due to the former providing better results for long horizons and the latter for short horizons.

[6] Note that apart from the literature on state-space models (see, e.g., Seong et al., 2013), few papers have considered non-stationary mixed-frequency variables. Pons and Sansó (2005) discuss the effects of the aforemen-

Chapter 2 of this thesis, jointly written with Alain Hecq and Jean-Pierre Urbain, presents a first step in this direction by proposing a mixed-frequency error correction model (ECM) for possibly cointegrated time series sampled at varying frequencies. More specifically, the forecasting performance of these ECM-MIDAS models is compared to models that omit the presence of a long-run relationship and to models obtained after temporally aggregating the high-frequency variable.[7] Indeed, such an analysis can be important for fore- or nowcasting business cycle indicators, because the gain of including the cointegrating relationship is apparent only for short horizons (Clements and Hendry, 1998). As far as the composition of the long-run term, i.e., *the dynamic mixed-frequency cointegrating relationship*, is concerned, we allow not only the end-of-period observation of the high-frequency variable to enter, but permit the 'neighboring' high-frequency observations to do so instead. While the property of cointegration is obviously not affected by this choice, it has an impact on the short-run dynamics terms of the ECM-MIDAS model at the model representation level.

It is found that ignoring a long-run term in the presence of cointegration lowers the forecasting performance of all approaches under consideration. Furthermore, the choice of a 'neighboring' high-frequency observation to enter the disequilibrium error has no effect on the forecast accuracy provided the structure of the short-run dynamics terms is adapted properly. If the latter requirement is not fulfilled, though, the forecasting performance of the respective approach may deteriorate. More specifically, while the inclusion of too few short-run variables leads to a worsened forecasting performances for all approaches (dynamic misspecification), over-specification of the short-run dynamics does not affect our mixed-frequency ECM models due to the data-driven weight determination underlying the MIDAS specification.

As discussed above, one of the most popular applications of mixed-frequency models is the improvement of macroeconomic forecasts. Aside from the feature of varying frequencies, macroeconomic indicators share another characteristic: Their figures get revised over time. Take quarterly GDP growth as an example. Even though the first figure corresponding to, say, the first quarter gets published in April, a new revised value appears in May, another one in June and so forth. Instead of making use of all of these revised figures, though, researchers often only consider the latest-available data points. Croushore and Stark (2001) argue that, in doing so, a data set is used that is different from the one researchers could have used in real time, and that this could have a drastic impact on the development of economic specifications.

In Chapter 3 of this thesis, also co-written by Alain Hecq and Jean-Pierre Urbain, the

tioned fixed temporal aggregation schemes on inference of cointegrating vectors. For $I(1)$ and $I(2)$ variables, it is shown that the aggregation scheme, which is best in terms of estimation precision, is not necessarily best in terms of hypothesis tests on the cointegrating vector. Furthermore, the asymptotic distributions of the respective coefficients in the various cointegrated systems are derived. Finally, the effects of different aggregation schemes on estimation are assessed via Monte Carlo studies.

[7] Independent from work presented in this thesis, Miller (2012) derives the asymptotic properties of the NLS estimator in more general mixed-frequency regressions with non-stationary, and possibly cointegrated, variables. So-called CoMIDAS regressions, which are essentially equivalent to our ECM-MIDAS regressions, then arise as special cases of these specifications. In another paper, Miller (2011) focuses on the efficient estimation of the cointegrating vector of a single-equation regression model involving a low-frequency regressand and high-frequency regressors.

analyses of real-time data sets and mixed-frequency variables are combined. In particular, by means of an application to U.S. growth, we extend the repeated observations forecasting approach of Stark and Croushore (2002) to an autoregressive distributed lag setting in which the regressors are of possibly higher frequencies than the dependent variable. Having access to regressors clears to road towards considering a wider range of models than the autoregressive ones in Stark and Croushore (2002). This leads to the question whether selection among a set of competing model specifications is sensitive to the data release chosen.

To investigate this issue, we compute data release-specific forecasts for a given calendar date and approximate the forecasts' distributions by continuous densities, which are subsequently combined along the set of models under consideration. The composition and evolvement of the resulting model-specific weights over time provide insights about the aforementioned research question, and allow the incorporation of revision process information into real-time studies. We find that, indeed, model selection is sensitive to the data release chosen, in the sense that the model ranking based on forecast accuracy measures obtained using latest-available data differs from the ranking implied by the aforementioned weights. In terms of incorporating revision process information into real-time studies, it turns out that forecast accuracy measures decrease with the forecast horizon and that forecasts made late in a quarter are more accurate than the ones made earlier.

Note that Chapters 2 and 3 deal with mixed-frequency variables in a single-equation dynamic regression model. If one were to pursue a system approach, such as a vector autoregressive (VAR) model, one would either temporally aggregate the high-frequency variable and work in a common low-frequency VAR or one would cast the model in state-space form, treating the high-frequency observations of the low-frequency variable as missing (Qian, 2013), and obtain a common high-frequency VAR. There has been no middle ground until the work of Ghysels (2012), which introduces an 'observation-driven' (Cox et al., 1981) mixed-frequency system by stacking the high- and low-frequency variables into a vector and formulating a VAR model in terms of this vector. The term 'observation-driven' refers to the feature that the mixed-frequency VAR does not involve any latent variables, but is formulated entirely in terms of observable data. Consequently, the model's impulse response functions are driven by observable shocks, which stands in contrast to 'parameter-driven' models that contain latent processes and, hence, latent shocks. Note that the latter is an undesirable feature as, e.g., policy shocks, are, of course, observable (Foroni and Marcellino, 2013).

Despite investigating a wide range of econometric issues (parsimony, shocks and estimation among others), the analysis in Ghysels (2012) is restricted to covariance-stationary variables. Similar to the single-equation case, merely transforming the variables to stationarity (by high- or low-frequency differencing) may not only neglect the presence of a cointegrating relationship, it may also lead to inefficiencies when cointegration between the high- and low-frequency variables is absent. Hence, an extension of mixed-frequency VAR models to the non-stationary, possibly cointegrated, case is of great importance. Chapter 4, also jointly written with Alain Hecq and Jean-Pierre Urbain, fills this gap in the literature by deriving a vector error correction model

(VECM) representation for non-stationary, possibly cointegrated, mixed-frequency variables.[8] Furthermore, these representations allow us to extend the common cyclical feature analysis of Engle and Kozicki (1993) or Vahid and Engle (1993) to a mixed-frequency setup.

We show that in the presence of non-stationary mixed-frequency variables, a researcher needs to account for two kinds of long-run relationships in a VAR, (i) a 'trivial' or prespecified set, which stems from the fact that the high-frequency variables are cointegrated with each other, and (ii) the long-run relationship between the low- and high-frequency variables. Incorporating the associated set of restrictions on the dynamics changes the set of variables that has to be used in the subsequent common feature analysis. It is found that our mixed-frequency VECM models perform quite well in practice and that the associated common feature test behaves very well (at least for the correct lag length). Also, common cycles that are hidden in a common low-frequency model may be detected using their mixed-frequency variants.

Being equipped with a mixed-frequency VAR or VECM sets the stage for an investigation of different sorts of issues, one of which is the notion of Granger causality. Due to the aforementioned VAR vector being obtained by stacking high- and low-frequency variables (corresponding to the same low-frequency time period), Granger causality is defined in terms of the low frequency. In Chapter 5 of this thesis, which is joint work with Alain Hecq, we analyze whether contemporaneous high-frequency observations help to explain, i.e., nowcast, the current value of the low-frequency variable and vice versa. In essence, we extend the notion of instantaneous causality (Lütkepohl, 2005, p.42) to the mixed-frequency setup and refer to it as *nowcasting causality*. We investigate the relationship between the latter and Granger causality in a mixed-frequency VAR and illustrate its impact on the parameters in a single-equation dynamic mixed-frequency regression model.

As far as Granger causality within mixed-frequency VAR models is concerned, Ghysels et al. (2013) develop a set of Granger causality tests that take advantage of the mixed-frequency nature of the variables involved. More precisely, they show that causal relationships (defined in the high frequency) are easier to recover using mixed frequencies compared to the approach based on temporal aggregation (low frequency). Furthermore, power improvements, asymptotically and in finite samples, are documented compared to conventional tests. The authors, however, make the implicit assumption that the frequency discrepancy between the variables involved is rather small, e.g., as in a quarter/month example. Chapter 6, co-written by Alain Hecq, relaxes this assumption and dicusses Granger causality testing in a mixed-frequency VAR, where the number of high-frequency observations per low-frequency period is large, e.g., as in a month/working day example.

[8] Seong et al. (2013) provide an example for a 'parameter-driven' VECM, i.e., one that contains latent processes, obtained by casting the model in state-space form. Independent from our work, Ghysels and Miller (2013) investigate cointegration tests for aggregated and mixed-frequency time series. In particular, starting from a common high-frequency VAR generating the data, the familiar trace test is defined and its limiting distribution derived. Subsequently, the authors analyze the impact of temporally aggregating either both variables (common low-frequency case) or only one variable (mixed-frequency case) on cointegration testing. They find that working in mixed frequencies is advantageous due to uncertainty surrounding the additional aggregation of the high-frequency variables.

The proliferation of parameters, resulting from the large number of parameters that need to be estimated, is addressed by two approaches, reduced rank restrictions and a Bayesian mixed-frequency VAR. The former rests on the assumption of a reduced rank within the submatrix of parameters corresponding to the lags of the high-frequency variables. The latter is implemented by properly extending the approach of Banbura et al. (2010) to our mixed-frequency setting. Indeed, due to stacking the high-frequency observations in the mixed-frequency VAR, their approach cannot be directly applied in the mixed-frequency case. Instead, a more careful investigation is required to properly incorporate prior beliefs into auxiliary dummy variables, i.e., into a prior distribution. Subsequently, these techniques are compared to the unrestricted mixed-frequency VAR and a common low-frequency VAR (based on temporal aggregation) in terms of their Granger (non-)causality testing behavior.

We confirm that the low-frequency VAR may lead to very poor results, due to Granger causality not being invariant to temporal aggregation (see Marcellino, 1999 or Breitung and Swanson, 2002), and that the unrestricted suffers from parameter proliferation for small sample sizes. While the performance of reduced rank restrictions depends crucially on how the observable high-frequency factors are computed (with overall best results for heterogeneous autoregressive type factors), the Bayesian mixed-frequency VAR approach yields size distortions when testing for causality in the direction from the low- to the high-frequency variables. As an empirical illustration, the link between uncertainty in financial markets and economic fluctuations is analyzed, whereby, in general, bi-directional Granger causality is detected.

Finally, Chapter 7 provides a short conclusion of the thesis.