*Innovativity in Modeling and Analytics Journal of Research*
**vol. 3, 2018, pp.72- 82** ***iMAJOR***
http://imajor.info/JDA ISSN 2534-9619

# Frequency transformation of the CPI for extending small datasets to big data

Borislava Vrigazova

Sofia University St.Kl. Ohridski, 125 Tsarigradsko Shosse Blvd., bl.3,
1113 Sofia, Bulgaria

bvrigazova@gmail.com

**Abstract.** Big data methods can be applied to large datasets. Economic datasets, however, provide a small number of observations as most economic statistics is available on a yearly basis. Relatively small number of economic indicators are presented on the basis of quarterly and monthly observations. Quarterly and monthly statistics is often not available for all countries. The question of how missing observations in quarterly and monthly statistics can be obtained is posed. The European Central Bank has long tried to solve this issue to obtain quarterly statistics for research as in [1]. Others as in [2] use mixed frequency data to perform their analysis. The aim of this research is to test various methods for frequency transformation of yearly data into quarterly data and analyze the difficulties in disaggregating yearly data. We perform our analysis in Python using the pandas and scipy packages. To verify our conclusions we compare the results of our research to official quarterly and monthly statistics provided by the World bank and Eurostat.

**Keywords:** frequency transformation, Python, big datasets

## 1 Introduction

Official statistics published by the World Bank, ECB, Eurostat, etc. is the main source of data for economic research. Most economic indicators, however, are published on a yearly basis, which can significantly decrease the size of the sample. With reduced sample size, the possibilities to apply big data to economic data are limited. This limitation has posed the problem of how the sample size can be increased in a reliable way. Traditional statistical approaches include cubic interpolation [3], linear interpolation [4], spline interpolation [5], polynomial interpolation [6] and quadratic interpolation [10]. Chin and Lo [7] and Danton [8] proposed another type of interpolation for disaggragation of time series often preferred by the official statistics [9]. In this paper, we test the proposed methods for disaggregation of the CPI and compare the forecasted quarterly data with official quarterly statistics. Our results suggest that the appropriateness of one method does not depend on the type of economic indicator (cpi, unemployment rate, etc) but rather on the country. This result implies that a change in the approach to temporal disaggragation of time series is needed. We believe that rather than applying one method to all countries in an economic indicator (e.g. the CPI) , the most appropriate disaggragation method for an economic indicator should be identified for each country separately. The resulting dataset will have the advantage that quarterly data could follow the patterns of the country’s economic development more closely, which preserve the quality of the quarterly data used in big data methods.

## 2 Literature review

Temporal disaggregation can be defined as the process of deriving high frequency data (e.g., monthly data) from low frequency data (e.g., annual data). Methods for temporal disaggragation can be classified into three groups. Statistical methods include interpolation methods like linear, quadratic, cubic, spline and polynomial interpolation [3,4,5,6 and 10]. Mathematical methods with application to socio-economic studies include the Chow-Lin method [11] for temporal disaggragation and Denton’s method [8] for benchmarking quarterly data from yearly data. The Chow-Lin method [11] derives quarterly data from yearly data by preserving the short-term behaviour of economic and social data but the quarterly series differ from the annual figure. The Denton’s method [8] uses annual data as a benchmark to disaggragate quarterly series and achieves consistency among frequencies. Econometric models [12,13] involve deriving quarterly data based on multivariate estimation of yearly factors affecting the specific time series, e.g. the inflation rate. Depending on the input data, quarterly estimations vary.

The first group of disaggragation methods include statistical methods for interpolation. The most common is linear interpolation [4]. Typically, linear interpolation ,like other types of interpolation, is a method for filling missing values. Missing values are extracted on the basis of linear relationship between two known data points. Linear interpolation for temporal disaggragation extracts quarterly data from yearly data in the same way. Yearly data are decomposed into quarterly data by assuming quarters to be missing values, which are linearly connected with yearly data. The yearly value can be set as a starting or ending quarter. Although some yearly data exhibit upward or downward trend, their quarters may have jumps, which can be left undetected by the linear interpolation. Quadratic interpolation [10], on the other hand, assumes that the relationship between the two data points is quadratic or polynomial of order two. An advantage of quadratic interpolation is that it can approximate cyclical data better than the linear interpolation as the linear interpolation is a private case of the polynomial interpolation [6]. The most widely used polynomial interpolation method [6] for economic data is cubic interpolation [3]. Cubic interpolation [3] approximates best cyclical economic data as it provides continuity among forecasted datapoints that look like the sine and cosine functions. Unlike linear and quadratic interpolation, cubic interpolation requires 4 points to forecast the missing data. Depending on whether we use the first or the last point, the interpolated result may differ, which is a disadvantage of the method. An extension of the cubic interpolation is also present. The Hermite cubic interpolation [14] improves cubic interpolation as it provides a higher degree of continuity.

In economic research, spline interpolation[5] is another widely used type of interpolation as it captures the fluctuations in data better than the cubic and hermite interpolation. Spline interpolation can forecast missing data using splines of n order depending on the type of polynomial that describes the dataset best. Spline interpolation can also enhance visualization of data [15]. The problem of decomposing yearly data into quarterly data requires preserving the yearly trend in quarterly data but also capturing the quarterly fluctuations specific to the dataset. As methods [4,5,6 and 10] for filling missing data may fail to capture quarterly fluctuations, some researchers suggested that temporal decomposition should be made via statistical models of reality [1 and 2]. These models include forecasting quarterly values of an economic indicator based on the values of other economic indicators as in [1]. An important step in this approach is to filter the noise in the predictors by using the Kalman filter [16 and 17]. The filtered data can, then, be used for forecasting quarterly values. Data used for forecasting, though filtered, may not be ready for use. An appropriate macroeconometric model, which models the connections in the dataset, should be found. Depending on the variables, approapriate models for macroeconomic purposes can be, for example, the DSGE [18].

As the dataset that was extracted from temporal disaggragation has to be suitable for economic research, the researcher should be able to generate the missing quarterly data that follow the pattern of the yearly observations. If forecasting models are used to achieve the aim, the researcher becomes dependent on the size and quality of the available data for forecasting. To overcome this problem, researchers [12] have proposed the Mixed Frequency Model (MIDAs) [19] and mixed frequency VAR [20], which better capture the dependencies between yearly and quarterly observations and provide better forecast only on the basis of the frequency of one economic indicator. Additional regressors are not necessary. A drawback of the MIDAs technique is that the forecast can be of low quality in case of missing data or small size of the available sample. With the increase of the sample size and the decrease of unavailable data, the forecast improves.

Another method for temporal disaggragation is the Chow-Lin method [11]. The method includes both low frequency data on the economic indicator and other data on high frequency. The short term trends in data can be preserved in this way. Estimation can be done via one or more high frequency economic indicators. This method can be used separately or can be combined with the benchmarking method of Denton [8]. Denton's method uses weighted quadratic optimisation to achieve consistency between economic indicators at different frequency. Denton's method can use a single time series to deduct the data on another frequency or a combination of a time series and other economic indicators. Unlike standard interpolation methods [3,4,5,6 and 10], the Chow -Lin method and the Denton's method can be applied only to economic data. As they are designed to capture the characteristics of economic data, their forecast can be more accurate. Offilcial statistics like the ECB have used Chow-Lin's and Denton's methods [8 and 11] as a methodology for temporal disaggragation [9].

Despite various methods available for temporal disaggragation, there is no universal method applied in economics. In the next section we will provide the theoretical framework behind the disaggragation methods, which practical advantages and disadvantages we will evaluate in our research. As we show, each method can be suitable for the same time series belonging to different country and period. We show that the Denton-Cholette methods [8 and 11] are not always the best disaggragation techniques for economic data.

## 3 Theoretical Framework

Statistical methods, which we have compared, include interpolation methods like linear, quadratic, cubic, spline and polynomial interpolation [3,4,5,6 and 10]. We first review linear interpolation as the simplest technique.

**Linear interpolation** is a private case of polynomial interpolation. It can be expressed as:

$$f(x_q) = y_q = \frac{y_{y1} - y_{y0}}{x_{y1} - x_{y0}} = (x_q - x_{y0}) + y_{y0} \qquad (1)$$

Linear interpolation represents a straight line, which passes through ($x_{y0}$;$y_{y0}$) and ($x_{y1}$;$y_{y1}$). By solving equation 1 using the known coordinates ($x_{y0}$;$y_{y0}$) and ($x_{y1}$;$y_{y1}$),the unknown ($x_q$,$y_q$) can be found. In the case of temporal decomposition, the known coordinates ($x_{y0}$;$y_{y0}$) and ($x_{y1}$;$y_{y1}$) are the yearly observations for two years, while the unknown point ($x_q$,$y_q$) is the quarterly data for quarters 2, 3 and 4. The values for each first quarters are assumed to coincide with the yearly observation. Many software programs can set the yearly observation to be used as a value for the first or the last quarter. In our analysis we have used the yearly observation as a first quarter value. As quarterly data need to preserve the pattern f yearly data, linear interpolation is performed under the constraint that:

$$CPIyear1 = (x_{y0}; y_{y0}) = \frac{\sum_{i=2}^{4} (x_{qi}; y_{qi})}{4} \qquad (2)$$

The polynomial of order 1 should provide such a solution to eq. 1 that the averaged quarterly values should equal the value of the yearly observation. Note that the averaged quarterly vales do not mean that each quarter has the same weight. In the case of quarterly CPI the weight of each quarter is very close to 25% of the yearly data.

Quadratic interpolation is a polynomial of order two, which is based on three points rather than two as in linear interpolation. The quadratic function with coefficients a,b and c describes the forecasted values of the CPI index (equation 3):

$$y = ax^2 + bx + c \ . \qquad (3)$$

The polynomial interpolation can be expressed by a general form (equation 4):

$$f(x) = c_0 + c_1 x + c_2 x^2 + ... + c_{N-1} x^{N-1} \qquad (4)$$

When N = 4, the degree of the polynomial equals 3, which is the case of the cubic interpolation. A disadvantage of polynomial interpolation is the complex form as the degree of the polynomial increases. The more complex the form, the more computationally exhaustive the method is. As we increase the order of the polynomial, the more likely it is some of the interpolated value to be under- or over- estimated and be far away from the true points. To avoid this problem, spline interpolation can be used.

The spline interpolation is another class of interpolation techniques. Spline interpolation involves dividing the N data points into N-1 intervals. A particular function that best describes the interval can be found for each interval. This function is called a spline. The spline function $f_i(x)$ , where i is the number of the interval, is a polynomial of lower order, which sets the number of data points in the interval. The solution of the problem is finding the spline coefficients. In order to be able to find the spline coefficients several conditions must be fulfilled.

First, the spline function must go through the first and last point of the interval . Equation 5 shows the first condition for a solution to the spline.

$$f_i(x_i) = y_i\,,\;\; f_i(x_{i+1}) = y_i \qquad (5)$$

Second, the curve segments must have the same slope in the region they join together.

$$f_i^{'}(x_{i+1}) = f_{i+1}^{'}(x_{i+1})\,. \qquad (6)$$

The third condition is given by eq.7:

$$f_i^{''}(x_{i+1}) = f_{i+1}^{''}(x_{i+1})\;\;. \qquad (7)$$

It expresses continuity of the curve segments at the points where they join together.

As spline interpolation may not be appropriate for economic data, many central banks prefer the Denton-Cholette method [8]. The method is shown by eq.8:

$$\min_{y_t}\sum_{t=2}^{sN}\left(\frac{y_t}{p_t}-\frac{y_{t-1}}{p_{t-1}}\right)^2 \qquad \sum_{t\subset T} y_t = y_{0,T} \text{ for T} = 1...\text{N}\,. \qquad (8)$$

The Denton-Cholette method is a variation of the Denton's method described in [7]. The point is to find the value y, which is the closest to preliminary value p by solving the modified quadratic equation 8.

We have chosen the mean absolute error as a measure of how close the disaggragated values to the real series. Equation 9 shows how the men square error is calculated:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \text{ (9).}$$

## 4 Results

The task of disaggregating a time series into quarterly data is difficult as in reality researchers do not have access to the original data. Although there are various method for temporal disaggragation, the European central bank has used the Chin-Lo Denton Cholette's method for benchmarking and disaggragating economic time series. Our study, however, suggests that this method may not provide the most accurate approximation of quarterly data to existing data. We performed our analysis on the CPI index for the countries presented in Table 1.

**Table 1: List of countries included in our research by using the CPI**

| | | | |
|---|---|---|---|
| Australia | Estonia | Israel | the New Zealand |
| Austria | Finland | Italy | Poland |
| Belgium | France | Japan | Russia |
| Brazil | Germany | Korea | Portugal |
| Canada | Greece | Latvia | Slovak |
| Chile | Hungary | Lithuania | Slovenia |
| China | Iceland | Luxembourg | South Africa |
| Colombia | Indonesia | Mexico | |
| Czech republic | India | the Netherlands | |
| Denmark | Ireland | the Norway | |

**Source: Eurostat**

We disaggragated the CPI index using the methods in section 3 and compared the results with the real quarterly data as we wanted to find the method, which provides the closest estimation to real data. The Chin-Lo Denton Cholette's method is the widest used by the central banks but our results showed that it is the closest approximation to real data in only 84% of the cases in our dataset. We found this result intriguing as it shows that the economic differences of countries reflect the connection between yearly and quarterly data and when the disaggragation method cannot capture them, the mean absolute error increases. Moreover, in 16% of the countries the Chin-Lo Denton Cholette's method failed to minimize the MAE and the forecasted data were far away from real data. We consider this proportion to be big. Table 2 presents the countries in which the Chin-Lo Denton Cholette's method was not the best /resulted in big MAE/ for getting the quarterly CPI index from yearly observations.

**Table 2: Mean absolute error for the best methods for temporal disaggragation of the CPI**

| **time series** | **frequency transformation** | **method** | **mae** |
|---|---|---|---|
| *cpi for Canada* | *yearly to quarterly* | *linear interpolation* | *6.853452E-01* |
| | | spline interpolation order =2 | 7.729855E-01 |
| | | cubic spline | 7.375714E-01 |
| | | time interpolation | 6.866883E-01 |
| | | quadratic | 6.757296E-01 |
| | | cubic interpolation | 6.782228E-01 |
| | | Chin-Lo Denton Cholette | 1.965745E+01 |

| **time series** | **frequency transformation** | **method** | **mae** |
|---|---|---|---|
| cpi for Chile | *yearly to quarterly* | *linear interpolation* | 1.251758E+00 |
| | | spline interpolation order =2 | 1.348844E+00 |
| | | cubic spline | 1.351853E+00 |
| | | time interpoaltion | 1.253816E+00 |
| | | quadratic | 1.244986E+00 |
| | | cubic interpolation | 1.243304E+00 |
| | | Chin-Lo Denton Cholette | 1.879968E+01 |

| **time series** | **frequency transformation** | **method** | **mae** |
|---|---|---|---|
| cpi for China | *yearly to quarterly* | *linear interpolation* | 9.396126E-01 |
| | | spline interpolation order =2 | 1.029495E+00 |
| | | cubic spline | 1.034237E+00 |
| | | time interpoaltion | 9.408870E-01 |
| | | quadratic | 9.142990E-01 |
| | | cubic interpolation | 9.149094E-01 |
| | | Chin-Lo Denton Cholette | 1.365892E+01 |

| **time series** | **frequency transformation** | **method** | **mae** |
|---|---|---|---|
| cpi for Lithuania | *yearly to quarterly* | *linear interpolation* | 9.829271E-01 |
| | | spline interpolation order =2 | 1.197832E+00 |
| | | cubic spline | 1.180090E+00 |
| | | time interpolation | 9.845870E-01 |
| | | quadratic | 9.862135E-01 |
| | | cubic interpolation | 9.939266E-01 |
| | | Chin-Lo Denton Cholette | 1.294114E+00 |

| **time series** | **frequency transformation** | **method** | **mae** |
|---|---|---|---|
| cpi for Slovenia | *yearly to quarterly* | *linear interpolation* | 1.131455E+00 |
| | | spline interpolation order =2 | 1.318215E+00 |
| | | cubic spline | 1.305004E+00 |
| | | time interpolation | 1.133605E+00 |
| | | quadratic | 1.131959E+00 |
| | | cubic interpolation | 1.135213E+00 |
| | | Chin-Lo Denton Cholette | 2.599127E+00 |

| **time series** | **frequency transformation** | **method** | **mae** |
|---|---|---|---|
| cpi for Japan | *yearly to quarterly* | *linear interpolation* | 3.833319E-01 |
| | | spline interpolation order =2 | 6.560754E-01 |
| | | cubic spline | 6.335892E-01 |
| | | time interpoaltion | 3.837201E-01 |
| | | quadratic | 3.775478E-01 |
| | | cubic interpolation | 3.789472E-01 |
| | | Chin-Lo Denton Cholette | 7.049463E+01 |

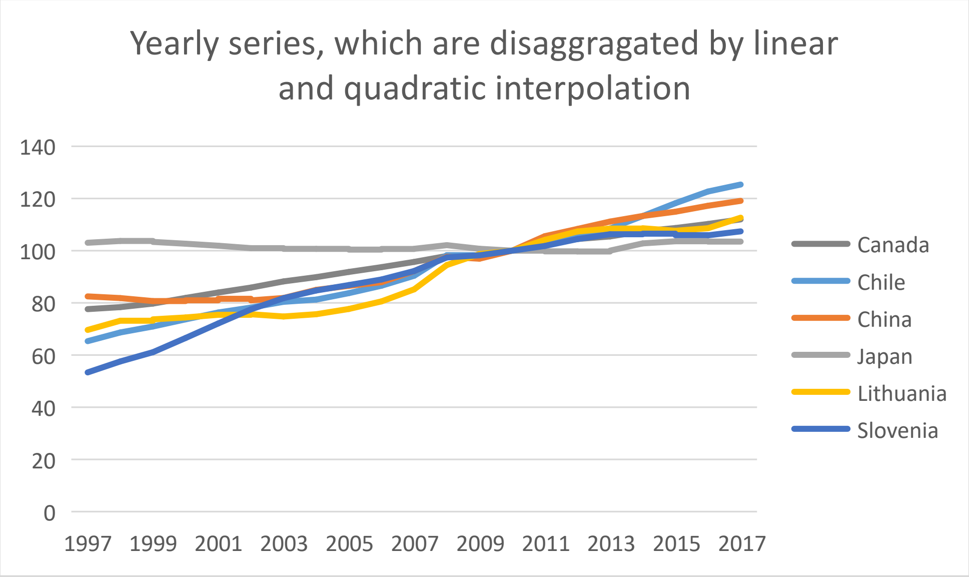
Source: authors’ calculations

As Table 2 shows the Chin-Lo Denton Cholette's method fails in 6 cases out of 37 countries. Temporal disaggregation for four of the countries is better performed by the quadratic interpolation and linear interpolation is the best in two of the cases. **Our study suggests two important findings.** First, one disaggregation method may not be appropriate for one indicator calculated for different countries due to economic and political differences. The same economic indicator may have to be disaggregated by a different method depending on the country. Second, one disaggregation method may be appropriate for application to a group of countries with similar characteristics. It should be noted that although the best method can be found on the basis of the MAE, the mean absolute error may still be large as it is the case of Slovenia. This poses the question about a benchmark of the mean absolute error , above which the method can be unreliable. So far, a benchmark for the error has not been identified.

Based on our results and the data, we can divide the sampled countries in three categories - countries for which the Chin-Lo Denton Cholette method is the best, countries for which the quadratic interpolation is the best and countries for which linear interpolation fits best. A surprising finding is that although there is a linear trend in the yearly series, linear interpolation is a good fit in only 5.4% of the cases.

Figure 1 shows the yearly CPI index for the six countries that are exception to Chin-Lo Denton Cholette's method.

**Fig.1 Six countries which do no follow the Chin-Lo Denton Cholette's method.**



**Source: Eurostat**

The figure shows that the CPI index in the six countries increase over time. However, the increase in the CPI for Lithuania and Slovenia seems to follow similar pattern. Each year the value of the CPI in these two countries has increased relatively smoothly. As a result, the linear interpolation disaggregated the quarterly series by minimizing the MAE. Although there is a linear trend in the other four series, there have been ups and downs in the series throughout the years.

The linear trend has not been smooth and as a result quadratic interpolation approximated the series best. For the rest of the 31 series the Chin-Lo Denton’s Cholette’s method proved to be the good fit. This is not surprising as there is a linear trend in the data but they are not smooth. **Our key finding suggests that there are three best methods for temporal disaggragation of economic time series depending on the smoothness of the linear trend in data.** If the linear trend is smooth, then linear interpolation will result in the smallest mean absolute error. If the linear trend exhibits small amount of curves, then the quadratic interpolation will minimize the MAE. The more curvy the trend, the more likely it is for the Chin-Lo Denton’s Cholette’s method to be the most appropriate method for temporal disaggragation. As the Chin-Lo Denton’s Cholette’s method is a modified version of the quadratic interpolation, the result is not surprising.

An important implication of our analysis is that when researchers try to extend their macroeconomic datasets, they should first try temporal disaggragation by the linear interpolation, the quadratic interpolation and the Chin-Lo Denton’s Cholette’s method. Our results suggest that other methods of temporal disaggragation like cubic interpolation and time interpolation may provide close estimates to the real data points, but still they are not close enough to be reliable. This is important conclusion as the values of the disaggragated series may affect the forecasting ability of the model.

## 5 Conclusion

As a conclusion, our study suggests that each economic time series for different countries should be disaggragated through different methods as the country can affect the quality of temporal disaggragation. In many cases, quadratic interpolation or the Chin-Lo Denton’s Cholette’s method can minimize the MAE and provide better approximation to the original quarterly series.

## 6 Acknowledgements

The author expresses his gratitude for the financial support to Sofia university under project 80-10-129/25.04.2018

## References

[1] J.Paredes, D. Pedregal, J. Perez, *A Quarterly Fiscal Database for the Euro Area Based on Intra-Annual Fiscal Information*. European Central Bank, Working Paper Series. 10.2139/ssrn.1537065, 2009.

[2] E. Ghysels, V. Kvedaras, V. Zemlys, *Mixed Frequency Data Sampling Regression Models: The R Package midasr*. Journal of Statistical Software. 72. 10.18637/jss.v072.i04, 2016.

[3] S. Durrleman, R. Simon, *Flexible regression models with cubic splines*. Statistics in Medicine. 1989;8:551–561. doi: 10.1002/sim.4780080504, 1989.

[4] Hazewinkel, Michiel, ed. , *Linear interpolation*, Encyclopedia of Mathematics, Springer Science+Business Media B.V. / Kluwer Academic Publishers, 2001.

[5] Nievergelt, Yves. UMAP: Module 718; Splines in Single and Multivariable Calculus. 1993. Lexington, MA: COMAP

[6] M. Gasca, T. Sauer, *On the history of multivariate polynomial interpolation*, Journal of Computational and Applied Mathematics, 122(1–2), 23-35, 2000, ISSN 0377-0427.

[7] G. Chow, A. Lin, *Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series*. The Review of Economics and Statistics, 53(4), 372-375. doi:10.2307/1928739, 1971.

[8] F. T. Denton, *Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization*, Journal of the American Statistical Association, 66:99–102, 1971.

[9] T. Di Fonzo, M. Marini, *On the Extrapolation with the Denton Proportional Benchmarking Method*, 2012, IMF Working Paper, 2012.

[10] Conn A.R., Toint P.L. (1996) An Algorithm using Quadratic Interpolation for Unconstrained Derivative Free Optimization. In: Di Pillo G., Giannessi F. (eds) Nonlinear Optimization and Applications. Springer, Boston, MA.

[11] G. C. Chow, A.-L. Lin, *Best linear unbiased interpolation, distribution, and extrapolation of time series by related series*, The Review of Economics and Statistics, 53(4):372–375, 1971.

[12] Andreou, E., E. Ghysels and E. Kourtellos, *Regression Models with Mixed Sampling Frequencies*, Journal of Econometrics, 158, 246-261, 2010.

[13] Asimakopoulos, S., J. Paredes, and T. Warmedinger (2013). Forecasting fiscal time series using mixed frequency data. Working Paper Series 1550, European Central Bank.

[14] Shahmorad with M. Zeinali, S & Mirnia, Kamal. (2013). Hermite and piecewise cubic Hermite interpolation of fuzzy data. Journal of Intelligent and Fuzzy Systems.

[15] M. Sarfraz, Malik Zawwar Hussain, *Data visualization using rational spline interpolation*, Journal of Computational and Applied Mathematics, 189(1-2), 513-525, 2006, ISSN 0377-0427.

[16] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Journal of Basic Engineering, 82, 35–45, 1960.

[17] Kalman, R. E and R. S. Bucy, *New Results in Linear Filtering and Prediction Theory*, 83, 95-108, 1961.

[18] L. Forni, L. Monteforte, L. Sessa, *The general equilibrium effects of fiscal policy: estimates for the euro area*, Journal of Public Economics, 93, 559-585, 2009.

[19] E. Ghysels, A. Sinko, R. Valkanov, *MIDAS Regressions: Further Results and New Directions*, Econometric Reviews, 26, 53-90, 2007.

[20] V. Kuzin, M. Marcellino,C. Schumacher, *MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area*, International Journal of Forecasting, Elsevier, 27(2), 529-542, 2011.