The 'tsdisagg2' Package

Jorge Alexandre Moura Alves Vieira
jorgealexandrevieira@gmail.com

October, 2016

Abstract

Time series disaggregation methods are used to disaggregate a low frequency time series to a higher frequency series with or without additional information contained in indicators. The 'tsdisagg2' is an R package, which implements the following disaggregation methods: Boot, Feibes and Lisman (both versions), Chow and Lin (both versions), Fernández and Litterman. This project was inspired by the CARMAX software, built in TSP by Santos Silva.

1 Time series disaggregation

The temporal disaggregation methods have evolved over the years. In 1962, Friedman suggested that it is possible to represent a time series as a linear process based on periodic observations [10]. Later, several changes occurred to the Friedman approach. On this paper, only the following approaches will be considered: Boot, Feibes and Lisman [2] in 1967 considered approaches based on univariate linear regression models (without the aid of associated series); Chow and Lin [4] in 1971, Fernández [8] in 1981 and Litterman [12] in 1983 considered approaches based on multivariate linear regression models (with the aid of associated series).

Let $\mathbf{Y}_{(N\times 1)}$ be the vector of the observed N annual values. Thus, using this methodology, it is intended to estimate $\mathbf{y}_{(n\times 1)}$, the n periodic values, so that $\mathbf{Y} = C\mathbf{y}$ where C is an aggregation matrix. This matrix is vital for the definition of the disaggregation techique to use: interpolation, distribution or extrapolation. The techinque depends on the type of variable that is available. Interpolation for stock variables and distribution for flow variables. It is noted that $n \geq sN$, wherein s is the frequency of observations (if $s = 4$, the observations are quarterly) . If $n > sN$, it is a case of extrapolation.

The flow variables may comply with one of the following restrictions:

– the sum of sub-periodic values is equal to the value of respective period

observation

$$Y_i = \sum_{k=1}^{s} y_{i,k}, \quad i = 1, ..., N, \quad k = 1, ..., s; \tag{1}$$

– the value observed on each period is the average of sub-periods values

$$Y_i = \frac{1}{s} \sum_{k=1}^{s} y_{i,k}, \quad i = 1, ..., N, \quad k = 1, ..., s. \tag{2}$$

The stock variables may comply with one of the following restrictions:

– the periodic total is equal to the value of the first sub-period

$$Y_i = y_{i,1} \quad i = 1, ..., N; \tag{3}$$

– the periodic total is equal to the value of the last sub-period

$$Y_i = y_{i,s}, \quad i = 1, ..., N. \tag{4}$$

Let $C_{(N \times n)}$ be the aggregation matrix which converts sub-periodic observed values into periodic values, therefore:

$$C = I_N \otimes c,$$

where $\otimes$ is the Kronecker product and $c_{(1 \times s)}$ is the aggregation vector. For a flow variable (case of distribution) that complies to restriction (1), the aggregation vector is given by $c = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}$. If the variable complies to restriction (2), then $c = \begin{bmatrix} \frac{1}{s} & \cdots & \frac{1}{s} \end{bmatrix}$. For a stock variable (case of interpolation) that complies to restriction (3), $c = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$. If the variable complies to (4), then $c = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}$.

It is possible to represent a time series as a linear process based on periodic observations, such as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{5}$$

where $\mathbf{y}_{(n \times 1)}$ is the vector of values to be estimated, $\mathbf{X}_{(n \times p)}$ is the matrix of observations of the available p indicators, $\beta_{(p \times 1)}$ is the vector of coefficients and $\epsilon_{(n \times 1)}$ is a residual variable with mean zero and variance-covariance matrix given by $\Sigma_{(n \times n)} = \sigma^2 \Omega$, where $\Omega_{(n \times n)}$ is the correlation matrix of ϵ. The model (5) is known as *high frequency model*.

As a consequence of the condition $\mathbf{Y} = C\mathbf{y}$, the periodic variable, $\mathbf{Y}$, can be written as

$$\mathbf{Y} = \dot{\mathbf{X}}\beta + \xi, \tag{6}$$

where $\dot{\mathbf{X}}_{(N\times p)} = C\mathbf{X}$ is the aggregated matrix of the available p indicators, $\xi_{(N\times 1)} = C\epsilon$ is a vector of random errors with variance-covariance matrix $\sigma_\epsilon^2 W$, with $W_{(N\times N)} = C\Omega C^T$. The model (6) is known as *low frequency model*.

The pioneers of time series disaggregation methodologies based on regression models were Chow and Lin (1971) [4] and since then, their methodology remains the most widely used. The developments made to their methodology focus on the autocorrelation pattern of errors generated by the model.

1.1 Methods

At an early stage of their research, Chow and Lin (1971) [4] found that residuals of the annual values are distributed evenly through the periodic observations, that is, they considered that residuals are stationary over time. In this case,

$$\Omega = I_n, \tag{7}$$

Later, Chow and Lin [4] suggested that the residuals follow an autoregressive process $AR(1)$, $\epsilon_t = \rho\epsilon_{t-1} + \delta_t$, with $t = 1, ..., n$, where δ_t is a white noise process and $|\rho| < 1$. Assuming that $\epsilon_0 = 0$,

$$\Omega = [(I_n + \rho L)^T (I_n + \rho L)]^{-1}, \tag{8}$$

where $L_{(n\times n)}$ is the auxiliar matrix

$$L = -1\times \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{9}$$

It should be noted that the last model requires the estimation of a new parameter, ρ.

Fernández (1981) [8] suggested a new method, similar to the aforementioned, which simplified the process of disaggregation. However, this simplicity is achieved through a strong assumption: $\rho = 1$. The author suggested that ϵ follows a *random walk* process, $\epsilon_t = \epsilon_{t-1} + \delta_t$, with $t = 1, ..., n$, where δ_t is a white noise process. This option simplifies the problem because there is no need to estimate the autocorrelation parameter. In this case,

$$\Omega = [D^T D]^{-1}, \tag{10}$$

with

$$D = I_n + L, \tag{11}$$

where L is the auxiliar matrix presented on (9).

Litterman (1983) [12] suggested a different approach, stating that the residuals are distributed through a *random walk-Markov* process. Thus, for $t = 1, ..., n$, $\epsilon_t = \epsilon_{t-1} + u_t$, where u_t is an $AR(1)$ process, $u_t = \rho u_{t-1} + \delta_t$, δ_t is white noise and $|\rho| < 1$. Assuming that $\epsilon_0 = u_0 = 0$, then

$$\Omega = [D^T (I_n + \rho L)^T (I_n + \rho L) D]^{-1}, \tag{12}$$

where L and D are the matrixes presented on (9) and (11), respectively.

In general, it is preferred that the time series disaggregation is performed using the information contained in associated indicators. When indicators are not available, it is appropriate to use simpler methods such as those suggested by Boot, Feibes and Lisman (1967) [2]. These authors suggested two methods: the first differences and the second differences methods. In the first, the only covariate is a constant and $\rho = 0$, so

$$\Omega = [D^T D]^{-1}, \tag{13}$$

where D is the matrix presented on (11).
In the second method, which is a particular case of Litterman's method, the only covariates are a constant and the time trend and $\rho = 1$, therefore

$$\Omega = [D^T D^T D D]^{-1}, \tag{14}$$

where D is the matrix presented on (11).

1.2 Estimation

For the estimation of the high frequency series, there is a need to estimate the parameters β, σ_ϵ^2 and ρ (when required). These parameters can not be estimated based on model (5) because $\mathbf{y}$ is unknown. Thus, the estimation is performed based on model (6). The BLU estimators for β and σ_ϵ^2 [1, 9, 13] are given by

$$\hat{\beta} = \left[\dot{\mathbf{X}}^T W^{-1} \dot{\mathbf{X}}\right]^{-1} \dot{\mathbf{X}}^T W^{-1} \mathbf{Y} \tag{15}$$

and

$$\hat{\sigma}_\epsilon^2 = (N - p)^{-1} \left[\mathbf{Y} - \dot{\mathbf{X}}\hat{\beta}\right]^T W^{-1} \left[\mathbf{Y} - \dot{\mathbf{X}}\hat{\beta}\right]. \tag{16}$$

To perform statistical inference on β, it is considered, for a large sample, that $\hat{\beta} \sim NMV\left(E\left[\hat{\beta}\right], Var\left[\hat{\beta}\right]\right)$, where $E\left[\hat{\beta}\right] = \beta$ and $Var\left[\hat{\beta}\right] = \hat{\sigma}_\epsilon^2 \left[\dot{\mathbf{X}}^T W^{-1} \dot{\mathbf{X}}\right]^{-1}$. Considering those conditions, it is possible to test the null hypothesis, $H_0 : \beta = \beta^*$, through Wald z-test [3, 16], where the statistic, $\mathcal{Z}_w$, is given by

$$\mathcal{Z}_w = \frac{\hat{\beta} - \beta^*}{\sqrt{Var\left[\hat{\beta}\right]}} \sim NMV\left(\mathbf{0}, \mathbf{1}\right). \tag{17}$$

The estimation of β and σ_ϵ^2 are based on the assumption that W is known. However, frequently, this assumption is not verified. Thus, the previous estimation of ρ is required. For the estimation of ρ, it can be performed a grid search for values within the interval $(-1, 1)$ with increments of 0.01. The optimal value is the one that maximizes the Objective Function (obtained from Concentrated Log-Likelihood function [7])

$$\Psi(\rho|\mathbf{Y}) = -i_\Psi \log\left|\hat{W}(\rho)\right| - N\log\left\{\left[\mathbf{Y} - \dot{\mathbf{X}}\hat{\beta}(\rho)\right]^T \hat{W}^{-1}(\rho)\left[\mathbf{Y} - \dot{\mathbf{X}}\hat{\beta}(\rho)\right]\right\}, \tag{18}$$

where

$$i_\Psi = \begin{cases} 1, & \text{for a Maximum Likelihood estimation;} \\ 0, & \text{for a Generalized Least Squares estimation.} \end{cases}$$

Recapping, with $\mathbf{Y}$ and $\mathbf{X}$ available, the estimation of β and ρ is possible. Being in possession of $\hat{\beta}$ and $\hat{\rho}$, makes possible the estimation of ξ, such that $\hat{\xi} = \mathbf{Y} - \dot{\mathbf{X}}\hat{\beta}$. Consequently, the estimator $\hat{\mathbf{y}}$ is obtained by correcting the naive estimate of $\mathbf{y}$, the linear combination $\mathbf{X}\hat{\beta}$, by distributing the residual low frequency estimates, $\hat{\xi}$, through that same estimate. This distribution is performed through the matrix

$$G = \Omega C^T W^{-1},$$

known as *Gain Projection Matrix* [11]. Thus, the estimator for $\mathbf{y}$ is given by

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} + G\hat{\xi} \\ &= \mathbf{X}\hat{\beta} + \Omega C^T W^{-1}\left[\mathbf{Y} - \dot{\mathbf{X}}\hat{\beta}\right]. \end{aligned} \tag{19}$$

The structure of $\hat{\mathbf{y}}$ ensures the important property of consistency $\mathbf{Y} = C\hat{\mathbf{y}}$.

2 About the 'tsdisagg2' package

The 'tsdisagg2' package was built in R [14] and was inspired by the CARMAX software, built in TSP, by Santos Silva. His software was originally built in 1996, as a project for INE-Portugal.

The 'tsdisagg2' package has only one available function, the `tsdisagg2` function. This function implements the following time series disaggregation methods: Boot, Feibes and Lisman (both versions), Chow and Lin (both versions), Fernández and Litterman.

2.1 Arguments

The role of each argument is described on Table 1:

Table 1: Arguments from 'tsdisagg2' function.

Arguments	Description
`y`	A *data.frame*, *matrix*, *list* or *vector* object with low frequency data.
`x`	A *data.frame*, *matrix*, *list* or *vector* object with high frequency data.
`da`	First year considered on low frequency data.
`dz`	Last year considered on low frequency data.
`s`	Frequency of observations; 3, 4 or 12; Default: `s=4`.
`method`	Set disaggregation method; 0 or 1; `'bfl1'`, `'bfl2'`, `'cl1'`, `'cl2'`, `'f'` or `'l'`); Default: `method='cl1'`
`c`	Model will be estimated with a constant; 0 or 1; Default: `c=0`.
`type`	Type of restriction; `'first'`, `'last'`, `'sum'` or `'average'`; Default: `type='sum'`.
`rho`	Sets a value for ρ; Any value within the interval $(-1, 1)$; Default: `rho=0`.
`neg`	The grid search is performed for negative or positive values of ρ; 0 or 1; Default: `neg=0`.
`ML`	Maximum Likelihood or Generalized Least Squares ρ estimation; 0 or 1; Default: `ML=0`.
`plots`	Generates the plot of the estimated series and the plot with the values from Objective Function; 0 or 1; Default: `plots=0`.

The definition of the disaggregation method depends on the `method` argument. Possible options for this argument are listed on Table 2.

Table 2: Methods of disaggregation.

Method	Ω	`method`
Chow e Lin	(7)	`'cl1'`
Chow e Lin	(8)	`'cl2'`
Fernández	(10)	`'f'`
Litterman	(12)	`'l'`
Boot, Feibes e Lisman	(13)	`'bfl1'`
Boot, Feibes e Lisman	(14)	`'bfl2'`

2.2 A quick example

The following example shows how to use `tsdisagg2` function. Consider the low frequency data shown on Table 3:

Table 3: Low frequency series.

Period	**Variable** Y
1995	203.92
1996	118.86
1997	139.82
1998	216.44
1999	291.03
2000	435.35

Consider that are known two indicators, associated to the low frequency series, X_1 and X_2. The high frequency data is shown on Table 4:

Table 4: High frequency series.

Sub-period	**Variable** X_1	**Variable** X_2
Mar-95	4778.96	58.65
Jun-95	5495.70	56.50
Sep-95	5145.27	45.16
Dec-95	4902.02	43.61
Mar-96	5883.39	34.30
Jun-96	5841.93	21.66
Sep-96	6201.72	32.07
Dec-96	6249.94	30.83
Mar-97	6413.88	16.46
Jun-97	6382.15	26.81
Sep-97	6723.71	43.86
Dec-97	6885.18	62.69
Mar-98	6928.36	59.60
Jun-98	7350.60	63.92
Sep-98	7844.95	54.86
Dec-98	8681.39	38.07
Mar-99	8857.55	70.07
Jun-99	8520.86	70.06
Sep-99	8328.24	64.12
Dec-99	7750.11	86.78
Mar-00	9154.53	100.85
Jun-00	7662.17	123.35
Sep-00	8045.06	115.17
Dec-00	8250.93	95.98

Consider that Y is a flow variable which complies to restriction (1). It is intended to disaggregate the (tranformed) low frequency series using Chow and Lin's method (second version), resorting to the aid of those indicators. It is intended a Maximum Likelihood estimation for negative values of ρ. From high frequency series, it is possible to verify that observations were recorded with a quarterly frequency. In R, the disaggregation of Y can be done by:

```
R> library(tsdisagg2)
R> annual=c( 203.92, 118.86, ..., 435.35 )
R> X1=c( 4778.96, 5495.70, ..., 8250.93 )
R> X2=c( 58.65, 56.50, 45.16, ..., 95.98 )
R> object <- tsdisagg2( y=annual, x=cbind( X1, X2 ), da=1995, dz
```

```
=2000, method='cl2', ML=1, neg=1, plots=1 )
```

Consequently, the following outputs are generated:

✓ A title indicating the method of disaggregation (`METHOD`):

```
Chow and Lin (AR1 residuals)
```

✓ The parameter estimates (`PARAMETERS ESTIMATION`) and the final series estimates:

```
Maximum Likelihood 'rho' estimation:  -0.77
(Loglik:  -29.28063 )

Sigma GLS:  4.56602
Sigma OLS:  6.843793

(Smooth:  30.79795 )


Model coefficients:
        estimate standard error z-statistic        p-value
X1 -0.0001829363   0.0002817168  -0.6493623  5.161042e-01
X2  1.0230230188   0.0323632102  31.6106780 2.633642e-219


Estimated values (Y-hat):
 year period     y-hat
 1995      1  58.33331
 1995      2  57.07181
 1995      3  44.72631
 1995      4  43.78857
 1996      1  33.64482
  ...    ...    ...
 2000      3 114.52033
 2000      4  96.65684
```

Note: The `Loglik` label refers to the value of the Log-Likelihood function

$$\ell(\beta, \sigma_\epsilon^2, \rho|\mathbf{Y}) = -\frac{N}{2}\log\left(2\pi\hat{\sigma}_\epsilon^2\right) - \frac{1}{2}\log\left|\hat{W}\right| - \frac{1}{2\hat{\sigma}_\epsilon^2}\left[\mathbf{Y} - \dot{\mathbf{X}}\hat{\beta}\right]^T \hat{W}^{-1}\left[\mathbf{Y} - \dot{\mathbf{X}}\hat{\beta}\right],$$

considering the optimal $\hat{\rho}$. Labels `Sigma OLS` and `Sigma GLS` are related to naive and robust estimates of σ_ϵ^2 (16), respectively. The `Smooth` value is an indicator of the smoothing degree of the estimated series given by

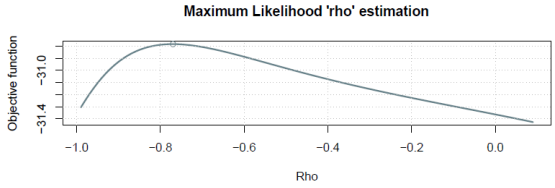
$$\hat{\mathbf{y}}^T D^T D^T D D \hat{\mathbf{y}} n^{-1}.$$

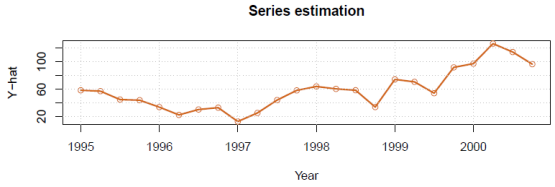
Lastly, in the `Model coefficients` table, are presented the β estimates and related results from Wald z-test (17).

✓ For each ρ tested, the respective Objective Function value (`OBJECTIVE FUNCTION`):

```
  rho     value
 0.09  -31.42703
  ...      ...
-0.97  -31.19837
-0.98  -31.25062
-0.99  -31.30589
```

✓ Objective Function and estimated series plots.





The aforementioned output is not interactable. Though, the interaction with generated results is still possible. As an example, to interact with the generated model residuals, it is advisable the following R code:

```
R> names( object )
R> object$RESIDUALS
```

References

[1] Aitken, A. C. (1934). "On least squares and linear combinations of observations", *Proceedings of the Royal Society of Edinburgh*, **55**, 42-48. URL `http://dx.doi.org/10.1017/S0370164600014346`

[2] Boot, J. C. G., Feibes, W., & Lisman, J. H. C. (1967). "Further methods of derivation of quarterly figures from annual data", *Applied Statistics*, **16**, 65-75. URL `http://dx.doi.org/10.2307/2985238`

[3] Cameron, A. C. & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*, Cambridge University Press. URL http://dx.doi.org/10.1017/CBO9780511811241

[4] Chow, G. & Lin, A. (1971). "Best linear unbiased interpolation, distribution and extrapolation of time series by Related Series", *The Review of Economics and Statistics*, **53**, 372-375. URL http://dx.doi.org/10.2307/1928739

[5] Denton, F.T. (1971). "Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization", *Journal of the American Statistical Association*, **66**, 99-102. URL http://dx.doi.org/10.1080/01621459.1971.10482227

[6] Di Fonzo, T. (2003). "Temporal disaggregation of economic time series: towards a dynamic extension", *Working Papers and Studies - Theme 1*, European Commission (Eurostat) - General Statistics. URL http://dx.doi.org/10.1007/BF02511587

[7] Dubin, R. A. (1988). "Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms". *The Review of Economics and Statistics*, **70**(3), 466-474. URL http://dx.doi.org/10.2307/1926785

[8] Fernández, R. B. (1981). "A methodologic note on the estimation of time series", *The Review of Economics and Statistics*, **63**, 471-476. URL http://dx.doi.org/10.2307/1924371

[9] Fomby, T. B., Hill, R. C., & Johnson, S. R. (2012). *Advanced Econometric Methods*. Springer Science & Business Media. URL http://dx.doi.org/10.1007/978-1-4419-8746-4

[10] Friedman, M. (1962). "The Interpolation of Time Series by Related Series", *Journal of American Statistical Association*, **57**, 729-757. URL http://dx.doi.org/10.1080/01621459.1962.10500812

[11] Goldberger, A. S. (1962). "Best Linear Unbiased Prediction", *Journal of the American Statistical Association*, **57**, 369-375. URL http://dx.doi.org/10.1080/01621459.1962.10480665

[12] Litterman, R. B. (1983). "A random walk, Markov model for the distribution of time series", *Journal of Business & Economics Statistics*, **1**, 169-173. URL http://dx.doi.org/10.2307/1391858

[13] Plackett, R. L. (1950). "Some Theorems in Least Squares", *Biometrika*, **37** (1-2), 149-157. URL http://dx.doi.org/10.2307/2332158

[14] R Core Team (2015). *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org

[15] Stromberg, K. R. (1981). *An Introduction to Classical Real Analysis.* Wadsworth, Belmont, CA. URL http://dx.doi.org/10.1090/chel/376

[16] Wald, A. (1943). "Test of Statistical Hypotheses Concerning Several Parameters when the Number of Observations Is Large", *Transactions of the American Mathematical Society*, **54**, 426-482. URL http://dx.doi.org/10.1090/S0002-9947-1943-0012401-3