

# Asymmetric Time Aggregation and its Potential Benefits for Forecasting Annual Data

Robert M. Kunst<sup>1</sup> and Philip Hans Franses<sup>2</sup>

Presented at the Royal Statistical Society Conference,  
Brighton, September 2010

---

<sup>1</sup>Institute for Advanced Studies, Vienna, Austria 1060, and University of Vienna;  
kunst@ihs.ac.at

<sup>2</sup>Erasmus University, Rotterdam, Netherlands 3000 DR



- 1 Introduction
- 2 Role-model examples
- 3 Time deformation
- 4 The algorithm
- 5 Monte Carlo evidence
- 6 Empirical application
- 7 Conclusions

# The aim of the research project

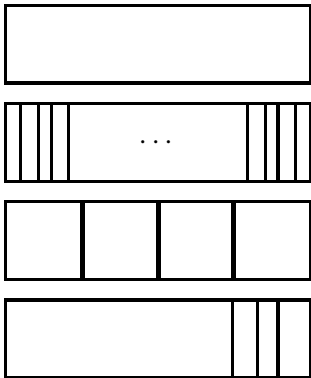
- The annual observation for a time-series variable that is available at a subannual frequency ( $S_1$  per year) should be predicted.
- The role-model case is a flow variable, such that the time aggregate is the sum (or a stock, provided the aggregate is an average).
- The seasonal information flow is heterogeneous across the year.



Christmas trees are extremely seasonal products. Most of them are sold some days before Christmas. Monthly sales data on January to October provide no information on next year's sales.



# Searching for the optimal time aggregation



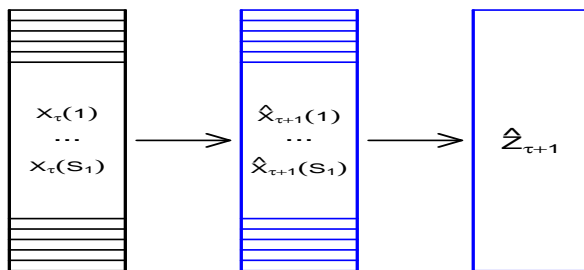
years: subannual information ignored

weeks ( $S_1 = 52$ ): use up degrees of freedom,  
some cells empty

quarters ( $S_2 = 4$ ): some cells almost empty

pseudo-quarters ( $S_2 = 4$ ): unequal time units,  
equal information spread

# Forecasting the next year



If the sample is large, a forecaster will estimate subannual values for the year  $\tau + 1$  as functions of the past, given all available observations until and including  $\tau$ , and aggregate the predicted values. In finite samples, estimates  $\hat{X}_{\tau+1}(w)$ ,  $w = 1, \dots, S_1$  and hence the forecast  $\hat{Z}_{\tau+1}$  may be poor.

# Evidence using Monte Carlo simulation

Large-sample asymptotics cannot reveal the strengths and weaknesses of the suggested regrouping scheme. We provide evidence by Monte Carlo simulation in finite samples and by application to empirical data.

A main impression is that asymmetric aggregation is most interesting for around 5 to 15 years of available data.

# Effects of regrouping in simple time-series models

Time disaggregation is not necessarily helpful in forecasting (see MAN, 2004). For example, if data are generated by a seasonal random walk  $X_{T,w} = X_{T-1,w} + \varepsilon_{T,w}$ ,  $E(X_T|X_{T-1})$  cannot be improved upon.

# Example 1

**Example 1:** Random walk on frequency  $S$ :

$$X_{\tau,w} = X_{\tau,w-1} + \varepsilon_{\tau,w}.$$

MSE for subannual-plus-aggregated forecast is  $\sigma_\varepsilon^2 \sum_{j=1}^S j^2$ , much less than for a forecast based on annual aggregates.

Optimal grouping into semesters ( $S_2 = 2$ ) is asymmetric: second pseudo-semester is just  $X_{\tau,S}$ , which yields an MSE equal to optimal forecast.

## Examples 2 and 3

**Example 2:** Seasonal random walk  $X_{T,w} = X_{T-1,w} + \varepsilon_{T,w}$ .

**Example 3:** Deterministic cycle plus noise  $X_{T,w} = \delta_w + \varepsilon_{T,w}$ .

In these examples, regrouping has no effect. Note, however, that dummy constants would be estimated in Example 3.

# What can be learned from the role-model examples?

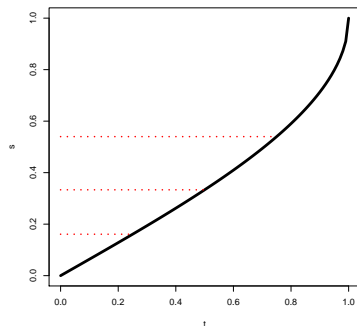
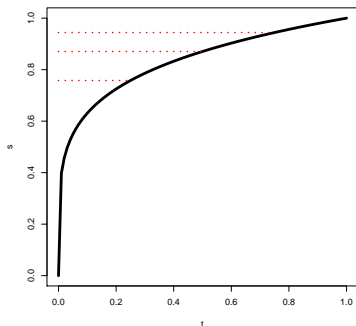
- Regrouping can be attractive in finite samples. In large samples, it can beat prediction based on annual data and naive aggregates at  $S_2$  but not fine observations at  $S_1$ .
- Regrouping has more effect if there is strong correlation within the year.
- Regrouping has more effect if there is heterogeneity in the generating laws across seasons.

# Seasonal time deformation

The concept of time deformation has been used much in finance (see, e.g., CLARK, 1973, GHYSELS, GOURIEROUX, AND JASIAK, 1995) and less often in business cycles (STOCK, 1987,1988) and in seasonal variation (JORDÀ AND MARCELLINO, 2004).

We use it for our generating processes. A highly correlated traditional AR process in economic time  $s$  is deformed, such that the information flow becomes seasonal in calendar time  $t$ .

# Time deformation functions



The functions  $s = t^\delta$  ('Box-Cox', left for  $\delta = 0.2$ ) and  $s = \frac{2}{\pi} \arcsin t$  ('arc-sine', right) imply that information is concentrated in the beginning or end of the year.

# The algorithm: the idea

Fine-frequency ( $S_1$ ) data are aggregated to coarse frequency  $S_2$ , such that the variation is spread equally among coarse time units.

This asymmetric time aggregation approximately reverts the time deformation in the observed fine-frequency data.

## The algorithm: the formula

Given estimates for the variance at seasonal units

$$\hat{\sigma}_w^2 = \frac{1}{T-1} \sum_{\tau=1}^T (X_{\tau,w} - \bar{X}_w)^2, w = 1, \dots, S_1,$$

the first artificial observation  $X_{T;1}$  accumulates all original data points  $X_{\tau,w}, w = 1, \dots, K$ , such that

$$K = \min\left\{k : \sum_{t=1}^{k-1} \sigma_w^2 + \frac{1}{2}\sigma_k^2 > \frac{\sigma^2}{S_2}\right\}.$$

The observation  $X_{T,K+1}$  at frequency  $S_1$  starts the second artificial observation  $X_{T;2}$ . This scheme is followed in analogy until all observations are regrouped.

# The simulation design

We generate 10,000 replications of AR processes in economic time  $s$  from

$$X_s = 0.99X_{s-1} + \varepsilon_s,$$

with Gaussian  $N(0, 1)$  errors, 250 observations per year. These are deformed using  $s = t^{0.2}$  and  $s = \frac{2}{\pi} \arcsin t$ .

We study two versions of observed ( $S_1$ ) and partially aggregated ( $S_2$ ) data: months and quarters ( $S_1 = 12, S_2 = 4$ ); and weeks and quarters ( $S_1 = 52, S_2 = 4$ ).

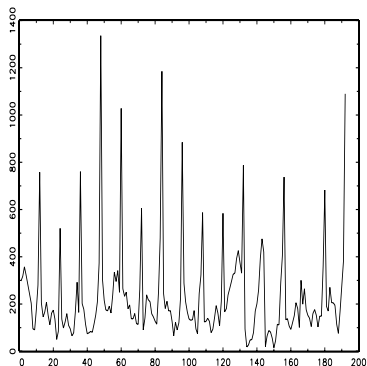
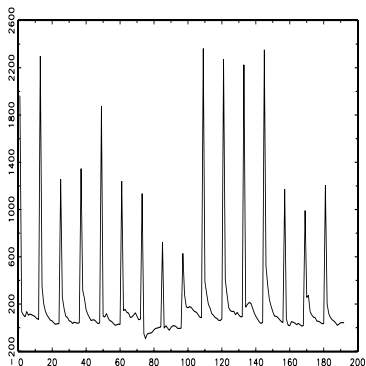
Time series are generated for 16 years. Assume  $\tau$  years are observed, and year  $\tau + 1$  is predicted.

# The considered forecasting models

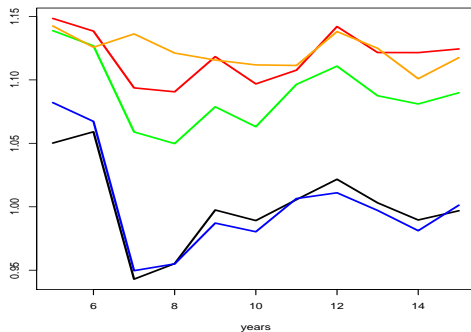
Four models are used for out-of-sample prediction:

- 1 AR model based on annual observations;
- 2 AR model based on fine frequency  $S_1$ ;
- 3 AR model based on coarse frequency  $S_2$  (quarters);
- 4 AR model based on regrouping algorithm and  $S_2$ .

AR lag orders are determined by AIC and BIC.



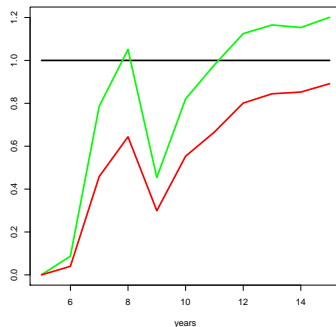
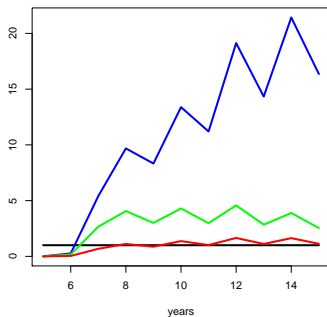
16 years of generated monthly data. Deformation functions are  $s = t^{0.2}$  and  $s = \frac{2}{\pi} \arcsin t$ .



Ratios of prediction MSE for regrouping algorithm to the model using traditional quarters. Curves represent  $\delta = 0.2$  (black),  $\delta = 0.3$  (blue),  $\delta = 0.4$  (green),  $\delta = 0.5$  (red), and  $\delta = 0.6$  (orange), with generating process based on  $t^\delta$ .

# What can be learned from the monthly simulation?

- If seasonality concentrates in the beginning of the year, all disaggregated models have similar performance; yearly aggregates are worse;
- If seasonality concentrates at the end of the year, seasonal 'gerrymandering' dominates; yearly aggregates work well for larger samples;
- If time deformation is weak (close to linear), regrouping is not rewarded.

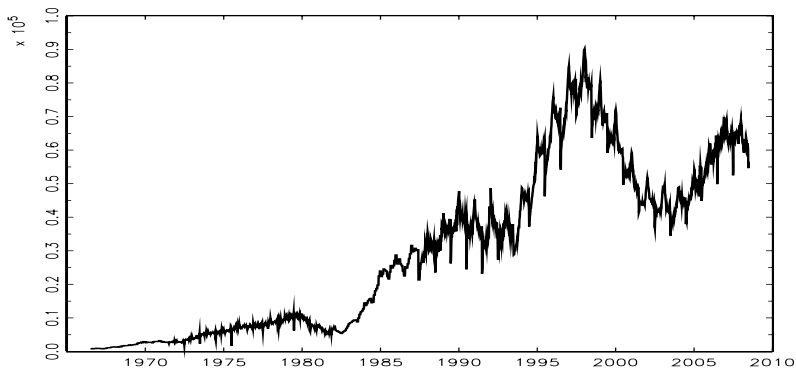


Ratios of prediction MSE for disaggregated prediction models to the annual model. Generating process uses  $\delta = 0.2$  in the left plot, and the arcsine model in the right plot. Black line represents the annual prediction, blue curve is for forecast based on weeks (left plot only), green for quarters, and red for regrouped pseudo-quarters.

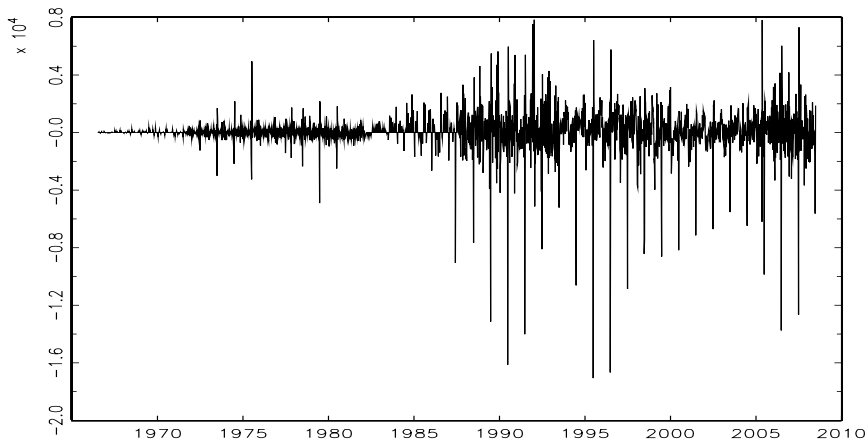
# What can be learned from the weekly simulation?

- Forecasts based on weeks are not competitive;
- Regrouping of weeks to pseudo-quarters dominates in those cases where it works for the previous experiment on months and quarters;
- For weeks and quarters, regrouped quarters always work better than calendar quarters.

# A weekly series of a stock variable



Weekly data for Randstad staffing services for the years 1967–2008.



First differences of the Randstad data.



## Forecasting annual values 1995–2008 of the Randstad data

	annual	weeks	quarters	pseudo-quarters
levels:				
MSE	8.7e+10	1.6e+11	2.7e+10	2.2e+10
# wins	2	3	4	5
ave. rank	2.86	3.07	2.14	1.93
differences:				
MSE	4.61e+7	4.99e+7	3.77e+7	3.82e+7
# wins	2	5	3	4
ave. rank	2.64	3.50	2.00	1.86

Note: MSE is the average squared error across the predicted years; ‘# wins’ is the number of cases where the respective model achieves the smallest error; ‘ave. rank’ is the average rank across all 14 cases.

# What can be learned from the empirical example?

- Forecasts based on regrouping perform slightly better than those based on calendar quarters (9 or 10 out of 14 cases);
- 14 cases of years are too few to admit a serious assessment of significance;
- Forecasts based on weeks or years models perform worse, which is in line with the simulations.

# General conclusions

- Regrouping ('seasonal gerrymandering') deserves attention if seasonal deformation is strong;
- if seasonal information is concentrated at the end of the year;
- if, say, 5 to 15 years of data are available.
- More designs will be investigated: for example, the influence of serial correlation in the generating process.

# Thank you for your attention