РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ
ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

РАНХиГС
экономический
факультет

РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И
ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРОГНОЗИРОВАНИЕ ИЕРАРХИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ

Касьянова Ксения

ЭО-15-01

Научный руководитель: Демешев Борис Борисович

Гипотеза:

- Используя модели учитывающие иерархическую структуру данных, мы сможем добиться значительного улучшения прогнозов для агрегированного временного ряда.

Цель:

- Улучшить прогноз аггрегированного ряда используя информацию из рядов второго и третьего уровня.

Задачи:

- Сбор данных с трехуровневой иерархической структурой
- Сравнение прогнозной силы моделей для агрегированного временного ряда, взвешенных прогнозов дизагрегрованных рядов и иерархической байесовской модели.
- Использование ближайшего по различным метрикам временного ряда в качестве регрессора
- Сравнение оптимальной комбинации прогнозов рядов второго уровня и сумм рядов третьего уровня, попавших в один кластер по различным метрикам

Данные:

Трехуровневая структура данных:

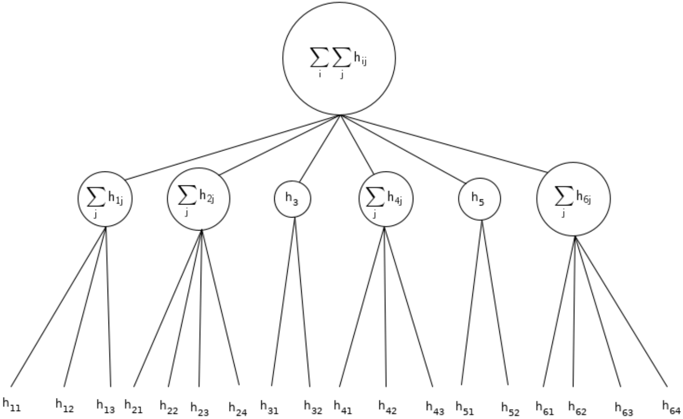


Рис.: Иерархическая структура временных рядов, необходимых для анализа

Данные:

Выбор подходящих наборов данных:

Для анализа были выбраны три набора данных с иерархической структурой с разной сезонностью:

- квартальные (ЕС):
  - ⇒ ВВП по 28 странам Европейского союза (включая Великобританию) в разбивке по 10 основным отраслям
  - ⇒ Данные собраны за период с 2000-Q1 по 2018-Q3
  - ⇒ Источник: Eurostat
- квартальные сезонно сглаженные (США):
  - ⇒ Данные по ВВП США (млн. долл., базовый год 2012) для каждого из 50 штатов с разбивкой по 21 основной отрасли
  - ⇒ Данные собраны за период с 2005-Q1 по 2018-Q2
  - ⇒ Источник: FRED
- месячные (РФ):
  - ⇒ Данные по смертности и рождаемости в каждом регионе, дающие в сумме естественный прирост населения РФ помесячно
  - ⇒ Данные собраны за период с 2006-01 по 2019-01
  - ⇒ Источник: ЕМИСС

Данные:

ЕС:

Квартальные данные по ЕС по секторам:
2000-01-01 по 2018-07-01

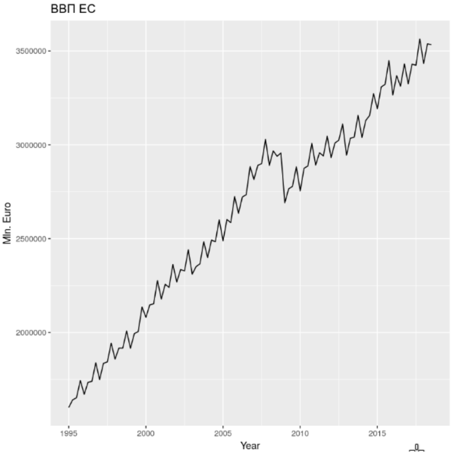


Рис.: ВВП ЕС28

Forecasting :

Aggregated TS:

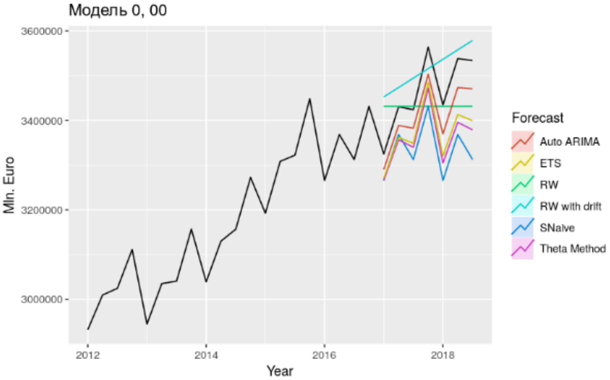
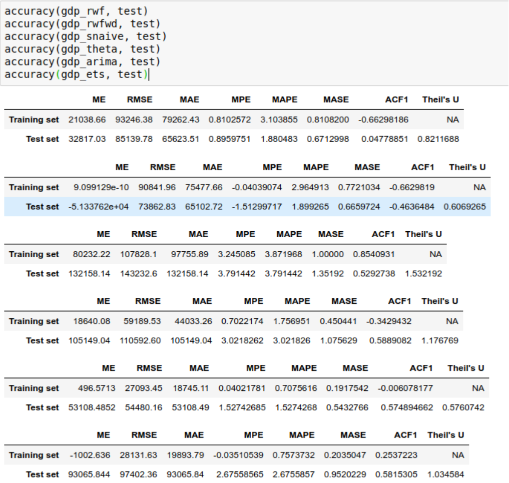


Рис.: Сравнение прогнозов

Forecasting :

Aggregated TS:



```
accuracy(gdp_rwf, test)
accuracy(gdp_rwfwd, test)
accuracy(gdp_snaive, test)
accuracy(gdp_theta, test)
accuracy(gdp_arima, test)
accuracy(gdp_ets, test)
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	21038.66	93246.38	79262.43	0.8102572	3.103855	0.8108200	-0.66298186	NA
Test set	32817.03	85139.78	65623.51	0.8959751	1.880483	0.6712998	0.04778851	0.8211688

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	9.099129e-10	90841.96	75477.66	-0.04039074	2.964913	0.7721034	-0.6629819	NA
Test set	-5.133762e+04	73862.83	65102.72	-1.51299717	1.899265	0.6659724	-0.4636484	0.6069265

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	80232.22	107828.1	97755.89	3.245085	3.871968	1.00000	0.8540931	NA
Test set	132158.14	143232.6	132158.14	3.791442	3.791442	1.35192	0.5292738	1.532192

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	18640.08	59189.53	44033.26	0.7022174	1.756951	0.450441	-0.3429432	NA
Test set	105149.04	110592.60	105149.04	3.0218262	3.021826	1.075629	0.5889082	1.176769

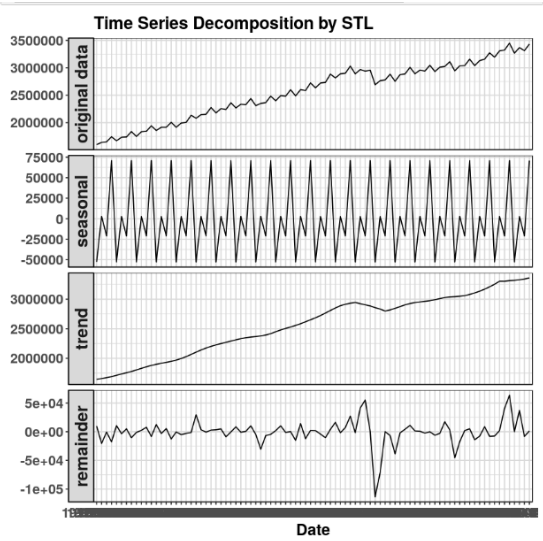
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	496.5713	27093.45	18745.11	0.04021781	0.7075616	0.1917542	-0.006078177	NA
Test set	53108.4852	54480.16	53108.49	1.52742685	1.5274268	0.5432766	0.574894662	0.5760742

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	-1002.636	28131.63	19893.79	-0.03510539	0.7573732	0.2035047	0.2537223	NA
Test set	93065.844	97402.36	93065.84	2.67558565	2.6755857	0.9520229	0.5815305	1.034584

Рис.: Сравнение прогнозов

Forecasting :

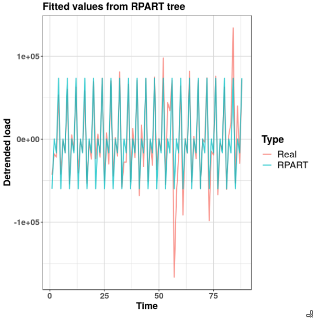
Aggregated TS:

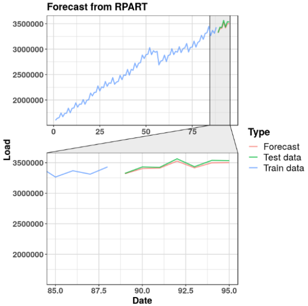


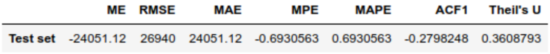
Forecasting :

Aggregated TS:

RPART, or CART (Classification and Regression Trees) is recursive partitioning type of binary tree for classification or regression tasks. It performs a search over all possible splits by maximizing an information measure of node impurity, selecting the covariate showing the best split.



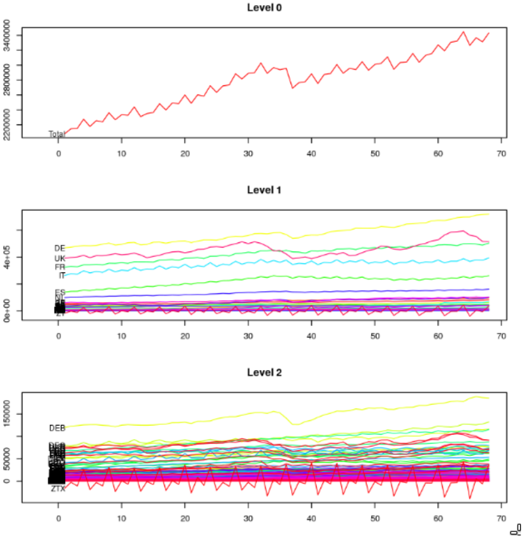




	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-24051.12	26940	24051.12	-0.6930563	0.6930563	-0.2798248	0.3608793

Forecasting :

Disaggregated TS



Grouped Time Series

Forecasting hierarchical time series

The assumption upon which many of these models are built on, is that by grouping series that behave in a similar way, the idiosyncratic errors within groups will tend to offset each other while the more relevant individual dynamics will be retained to be modelled.
Key idea: forecast reconciliation

- Ignore structural constraints and forecast every series of interest independently.
- Adjust forecasts to impose constraints.

Existing methods:

- Bottom-up
- Top-down
- Middle-out

An “optimal combination” approach can be advanced by proposing two new estimators based on WLS.
Both now implemented in the hts package

Grouped Time Series

Forecasting hierarchical time series

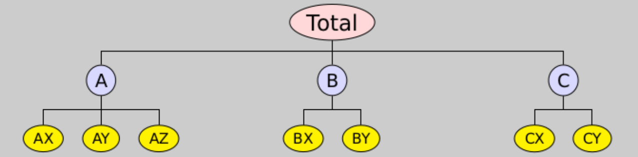


Рис.: Hierarchical structure

Y_t : observed aggregate of all series at time t.
$Y_{X,t}$: observation on series X at time t.

Grouped Time Series

Forecasting hierarchical time series

For the hierarchical structure we can write:

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix}$$

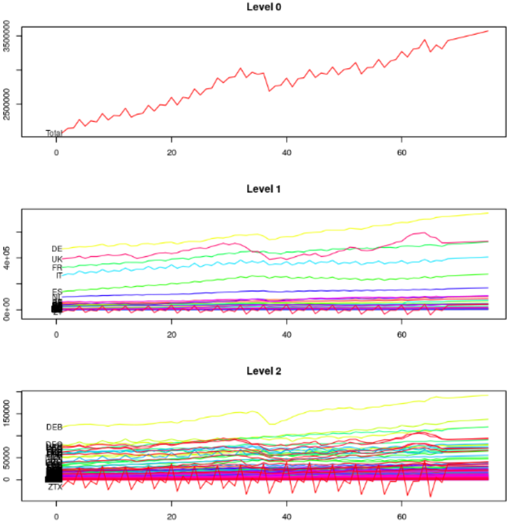
or in more compact notation, where y_t is an n-dimensional vector of all the observations in the hierarchy at time t, S is the summing matrix.

$$y_t = Sb_t$$

b_t : m-dimensional vector of all series at bottom level in time t.

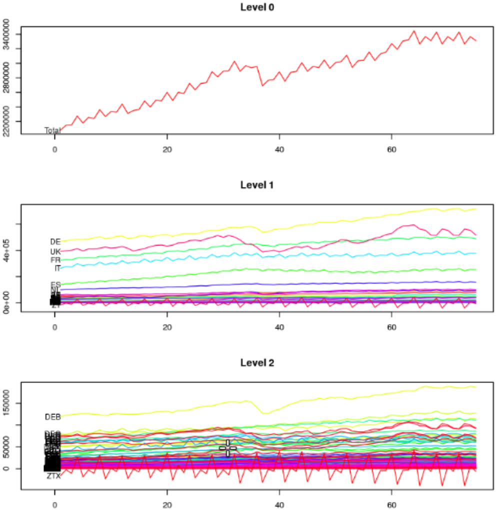
Grouped Time Series

RW with drift



Grouped Time Series

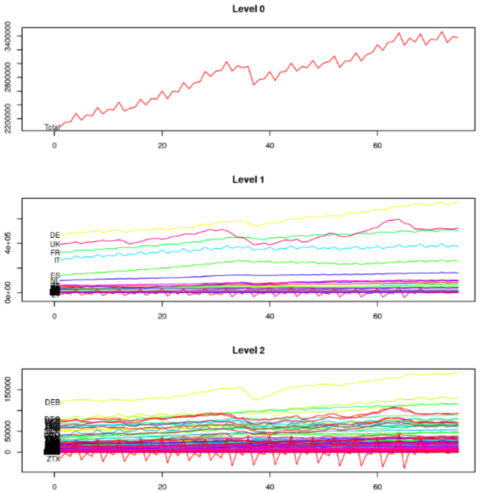
Snaive



Grouped Time Series

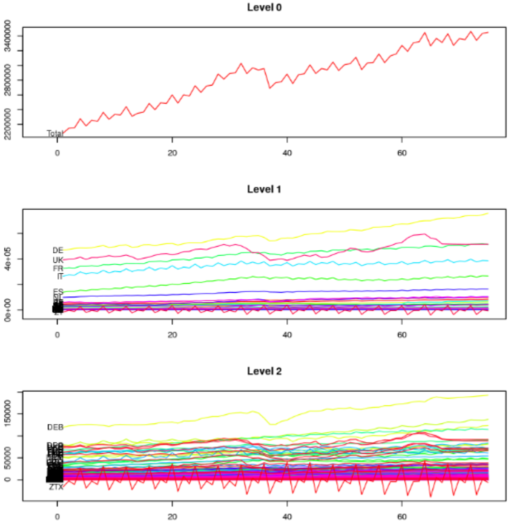
Theta

Equivalent to simple exponential smoothing with drift

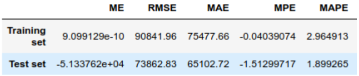


Grouped Time Series

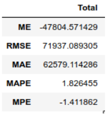
ARIMA



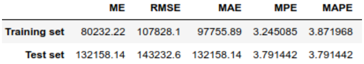
Grouped Time Series



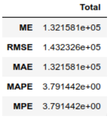
	ME	RMSE	MAE	MPE	MAPE
Training set	9.099129e-10	90841.96	75477.66	-0.04039074	2.964913
Test set	-5.133762e+04	73862.83	65102.72	-1.51299717	1.899265



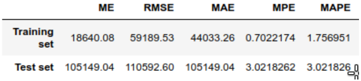
	Total
ME	-47804.571429
RMSE	71937.089305
MAE	62579.114286
MAPE	1.826455
MPE	-1.411862



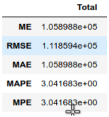
	ME	RMSE	MAE	MPE	MAPE
Training set	80232.22	107828.1	97755.89	3.245085	3.871968
Test set	132158.14	143232.6	132158.14	3.791442	3.791442



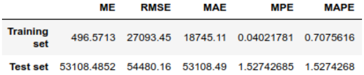
	Total
ME	1.321581e+05
RMSE	1.432326e+05
MAE	1.321581e+05
MAPE	3.791442e+00
MPE	3.791442e+00



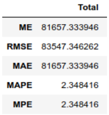
	ME	RMSE	MAE	MPE	MAPE
Training set	18640.08	59189.53	44033.26	0.7022174	1.756951
Test set	105149.04	110592.60	105149.04	3.0218262	3.021826



	Total
ME	1.058988e+05
RMSE	1.118594e+05
MAE	1.058988e+05
MAPE	3.041683e+00
MPE	3.041683e+00



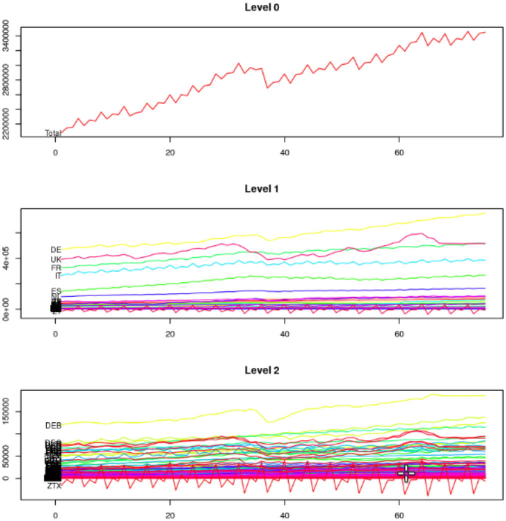
	ME	RMSE	MAE	MPE	MAPE
Training set	496.5713	27093.45	18745.11	0.04021781	0.7075616
Test set	53108.4852	54480.16	53108.49	1.52742685	1.5274268



	Total
ME	81657.333946
RMSE	83547.346262
MAE	81657.333946
MAPE	2.348416
MPE	2.348416

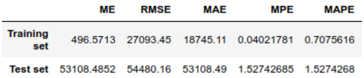
Grouped Time Series

ARIMAX

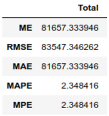


Grouped Time Series

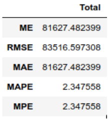
ARIMAX



	ME	RMSE	MAE	MPE	MAPE
Training set	496.5713	27093.45	18745.11	0.04021781	0.7075616
Test set	53108.4852	54480.16	53108.49	1.52742685	1.5274268



	Total
ME	81657.333946
RMSE	83547.346262
MAE	81657.333946
MAPE	2.348416
MPE	2.348416



	Total
ME	81627.482399
RMSE	83516.597308
MAE	81627.482399
MAPE	2.347558
MPE	2.347558

Grouped Time Series

Forecasting grouped time series

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{X,t} \\ y_{Y,t} \\ y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix}$$

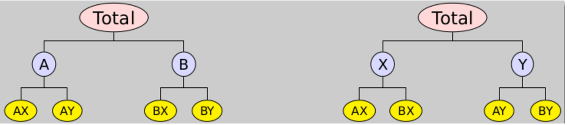


Рис.: Grouped structure

Grouped Time Series

Optimal forecasts

$$\hat{Y}_n(h) = S\beta_n(h) + e_h$$

$\hat{Y}_n(h)$ - a vector of initial h-step forecasts, made at time n, stacked in same order as Y_t.
$\beta_n(h) = E[B_{n+h}|Y_1, \dots, Y_n]$
e_h has zero mean and covariance Σ_h.

$$\tilde{Y}_n(h) = S\hat{\beta}_n(h) = SP\hat{Y}_n(h) = S(S'\Sigma_h^+ S)^{-1} S'\Sigma_h^+ \hat{Y}_n(h)$$

$\tilde{Y}_n(h)$ - revised forecasts, $\hat{Y}_n(h)$ - initial forecasts
Σ_h^+ - generalized inverse of Σ_h.
Optimal $P = (S'\Sigma_h^+ S)^{-1} S'\Sigma_h^+$
Problem: Σ_h hard to estimate.

Grouped Time Series

Approximate optimal forecasts

$$\tilde{Y}_n(h) = S(S'\Sigma_h^+ S)^{-1} S'\Sigma_h^+ \hat{Y}_n(h)$$

- OLS
  $e_{B,h}$ is the forecast error at bottom level.
  Assume $e_h \approx S e_{B,h}$ then $(S'\Sigma^+ S)^{-1} S'\Sigma^+ = (S'S)^{-1} S'$

$$\tilde{Y}_n(h) = S(S'S)^{-1} S'\hat{Y}_n(h)$$

- Rescaling
  Suppose we approximate Σ_h by its diagonal: $\Lambda = diag(\hat{\Sigma}_1)^{-1}$, which contain inverse one-step ahead in-sample forecast error variances.

$$\tilde{Y}_n(h) = S(S'\Lambda S)^{-1} S'\Lambda\hat{Y}_n(h)$$

- Averaging
  If the bottom level error series are approximately uncorrelated and have similar variances, then Λ is inversely proportional to the number of series contributing to each node: $\Lambda = diag(S \times 1)^{-1}$ is the inverse row sums of S

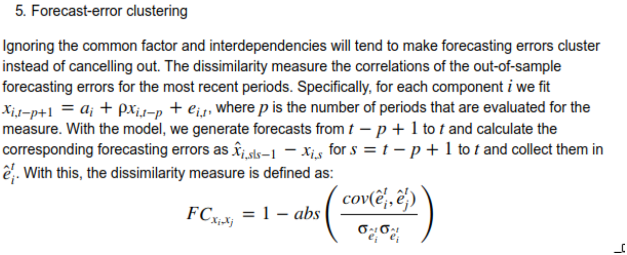
Grouped Time Series

Temporal hierarchies

Basic idea:

Forecast series at each available frequency. Optimally combine forecasts within the same year.

Forecast-error clustering

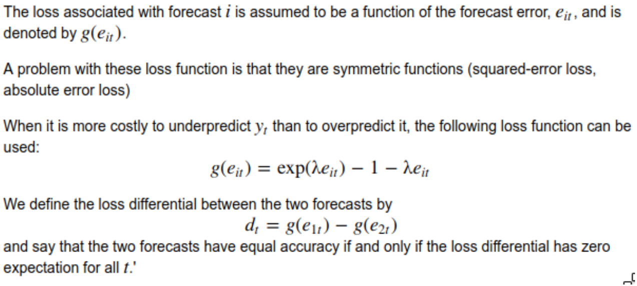


5. Forecast-error clustering

Ignoring the common factor and interdependencies will tend to make forecasting errors cluster instead of cancelling out. The dissimilarity measure the correlations of the out-of-sample forecasting errors for the most recent periods. Specifically, for each component i we fit $x_{i,t-p+1} = a_i + \rho x_{i,t-p} + e_{i,t}$, where p is the number of periods that are evaluated for the measure. With the model, we generate forecasts from $t - p + 1$ to t and calculate the corresponding forecasting errors as $\hat{x}_{i,s|s-1} - x_{i,s}$ for $s = t - p + 1$ to t and collect them in $\hat{e}_i^t$. With this, the dissimilarity measure is defined as:

$$FC_{x_i,x_j} = 1 - abs\left(\frac{cov(\hat{e}_i^t, \hat{e}_j^t)}{\sigma_{\hat{e}_i^t}\sigma_{\hat{e}_i^t}}\right)$$

The Diebold-Mariano Test



The loss associated with forecast i is assumed to be a function of the forecast error, e_{it}, and is denoted by $g(e_{it})$.

A problem with these loss function is that they are symmetric functions (squared-error loss, absolute error loss)

When it is more costly to underpredict y_t than to overpredict it, the following loss function can be used:

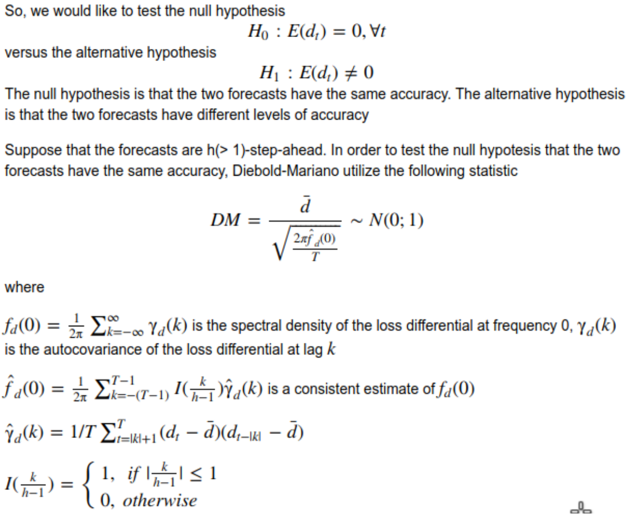
$$g(e_{it}) = \exp(\lambda e_{it}) - 1 - \lambda e_{it}$$

We define the loss differential between the two forecasts by

$$d_t = g(e_{1t}) - g(e_{2t})$$

and say that the two forecasts have equal accuracy if and only if the loss differential has zero expectation for all t.'

The Diebold-Mariano Test



So, we would like to test the null hypothesis

$$H_0 : E(d_t) = 0, \forall t$$

versus the alternative hypothesis

$$H_1 : E(d_t) \neq 0$$

The null hypothesis is that the two forecasts have the same accuracy. The alternative hypothesis is that the two forecasts have different levels of accuracy

Suppose that the forecasts are h(> 1)-step-ahead. In order to test the null hypotesis that the two forecasts have the same accuracy, Diebold-Mariano utilize the following statistic

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}} \sim N(0;1)$$

where

$f_d(0) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_d(k)$ is the spectral density of the loss differential at frequency 0, $\gamma_d(k)$ is the autocovariance of the loss differential at lag k

$\hat{f}_d(0) = \frac{1}{2\pi} \sum_{k=-(T-1)}^{T-1} I(\frac{k}{h-1})\hat{\gamma}_d(k)$ is a consistent estimate of $f_d(0)$

$$\hat{\gamma}_d(k) = 1/T \sum_{t=|k|+1}^{T} (d_t - \bar{d})(d_{t-|k|} - \bar{d})$$

$$I(\tfrac{k}{h-1}) = \begin{cases} 1, & if\ |\frac{k}{h-1}| \leq 1 \\ 0, & otherwise \end{cases}$$

The Diebold-Mariano Test

Harvey, Leybourne, and Newbold (1997) (HLN) suggest that improved small-sample properties can be obtained by:

1. making a bias correction to the DM test statistic, and
2. comparing the corrected statistic with a Student-t

distribution with (T-1) degrees of freedom, rather than the standard normal.

$$HLN = DM\sqrt{(n+1-2h+h(h-1))/n} \sim T(n-1)$$

- The Diebold-Mariano test should not be applied to situations where the competing forecasts are obtained using two nested models, since at the population level, if the null hypothesis of equal predictive accuracy is true, the forecast errors from the competing models are exactly the same and perfectly correlated, which means that the numerator and denominator of a Diebold-Mariano test are each limiting to zero as the estimation sample and prediction sample grow.
- However, when the size of the estimation sample remains finite, parameter estimates are prevented from reaching their probability limits and the Diebold-Mariano test remains asymptotically valid even for nested models, under some regularity assumptions (see Giacomini and White 2003).

Bayesian Methods

Forward modelling: with noise properties we can predict the Sampling Distribution (the probability for a general set of data

In easy cases, the effect of the prior is simple. As experiment gathers more data, the likelihood tends to get narrower, and the influence of the prior diminishes.

Rule of thumb: if changing your prior to another reasonable one changes the answers significantly, you need more data

Bayesian Methods

Why is the Bayesian Inference controversial?

1. Bayesians claim that the parameters are random so that their credible interval is a valid probability argument, though it also depends on the the prior, which is usually hard to obtain.
2. When the likelihood and prior is complicated, the inference has to rely on the MCMC sampling, which can be really slow in most of the real-world cases.
3. The biggest controversy about Bayesian inference is that you must quantify your prior knowledge about the question at hand. This makes it possible to actually influence your results, either accidentally or on purpose.

Bayesian Methods

Bayesian Modelling and Markov chain Monte Carlo

Bayesian method provides an intuitive way for us to fill the gaps left by small or incomplete data sets.
To calculate the Bayesian predictive distribution, $\pi(x|D)$, given some data D, we simply multiply the density function of the classical solution $\ell(D|x)$, with the density function produced by our prior knowledge $\pi(x)$. This is a direct application of Bayes' theorem. Unfortunately, this product will not integrate to one. To overcome this, we multiply the density function by a constant Z, which rescales the density so that it does integrate to one.
The resulting Bayesian distribution defined over the n-dimensional parameter space S is

$$\pi(x|D) = \frac{1}{Z}\pi(x)\ell(D|x)$$

$$Z = \int_S \pi(x)\ell(D|x)dx$$

In one dimension it is easy to use numerical quadrature to calculate Z.
However as the dimension becomes large, this method quickly becomes impractical. So we turn to a class of statistical algorithms known as Markov chain Monte Carlo (MCMC) methods, which can tackle these high dimensional parameter spaces.

Bayesian Methods

Bayesian Shrinkage

Shrinkage is implicit in Bayesian inference and penalized likelihood inference, and explicit in James–Stein-type inference. In contrast, simple types of maximum-likelihood and least-squares estimation procedures do not include shrinkage effects, although they can be used within shrinkage estimation schemes.

Bayesian Methods

Stein's paradox

Stein's paradox, in decision theory and estimation theory, is the phenomenon that when three or more parameters are estimated simultaneously, there exist combined estimators more accurate on average (that is, having lower expected mean squared error) than any method that handles the parameters separately.

The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better that the arithmetic average

Stein's paradox **modern generalization, the Bayesian hierarchical model**.

Bayesian hierarchical model can improve overall estimation accuracy, thereby improving our confidence in the assessment results, especially for standard compliance assessment of waters with **small sample sizes.**

Small Sample

Most of the inferential procedures available in the analysis of time series data are asymptotic

Although analytic small sample results are available in a few cases, there is currently, no widely applicable and easily accessible method that can be used to make small sample inferences.

Bootstrap

The bootstrap technique introduced by Efron (1979) could possibly be a potential alternative in estimation and inference from time series models in finite samples. However, in time series regressions, the standard bootstrap resampling method designed for independent and identically distributed (IID) errors is not applicable because in most situations the assumption of IID errors is violated.

The basic bootstrap approach consists of drawing repeated samples (with replacement). The simplest assumption for that method is that observations should be IID. But in time series models IID assumption is not satisfied. Thus the method needs to be modified.

BS for TS:

- Estimating Standard Errors: if “Small Sample Size” distribution is normal then we can get a BS distribution to Estimate SE (same as asymptotic distribution for SE)
- Confidence Interval statements: Using BS distribution to Estimate CI we can get different result for CI (from asymptotic distribution), for example, because of BS distribution skewness

Bootstrap

The Recursive BS for stationary AR(p) model

Consider AR(p) process:

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + e_t, e_t \sim N(0,\sigma^2)$$

We estimate coefficients with OLS and get:

$$(\hat{a}_1, \ldots, \hat{a}_p), \hat{e}_t$$

Define the centered and scaled residuals:

$$\tilde{e}_t = (\hat{e}_t - \frac{1}{n}\sum \hat{e}_t)\left(\frac{n}{n-p}\right)^{1/2}$$

Resample $\tilde{e}_t$ with replacement to get the BS residuals e_t^*
Construct the BS sample recursively using $y_t^* = y_t$:

$$y_t^* = \sum_{i=1}^{p} \hat{a}_i y_{t-i}^* + e_t^*$$

Bootstrap

A comparison of four different block bootstrap methods

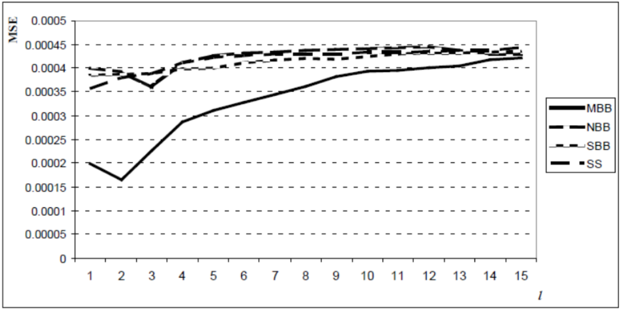


Figure 1: *Preliminary comparison of block bootstrap methods*

Рис.: MBB – Moving block bootstrap, NBB – Non-overlapping block bootstrap, SBB – Stationary block bootstrap, SS - Subsampling

Bootstrap

Bootstrapping vs. Posterior sampling

“Bootstrapping” is a a framework that aims to improve simple but approximate frequentist methods:

- Parametric bootstrap: Improve asymptotic behavior of estimates for a trusted model: reduce bias of estimates, provide more accurate coverage of confidence regions
- Nonparametric bootstrap: Provide results that are approximately accurate with weak modeling assumptions

Most common approach uses Monte Carlo to simulate an ensemble of data sets related to the observed one, and use them to recalibrate a simple method.

Parametric bootstrap has a step producing an ensemble of estimates that looks like a set of posterior samples. Can they be thought of this way?