MONASH University

MONASH BUSINESS SCHOOL

2017 Beijing Workshop on Forecasting

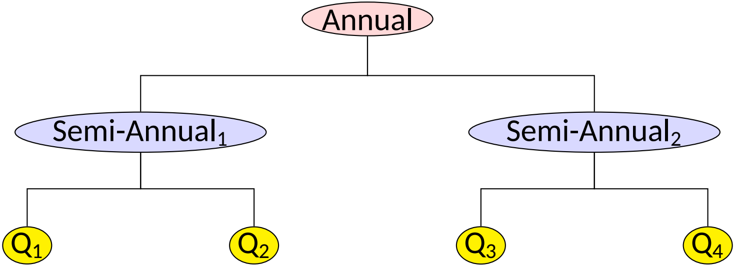
Probabilistic Hierarchical Forecasting

Rob J Hyndman

robjhyndman.com/beijing2017

Outline

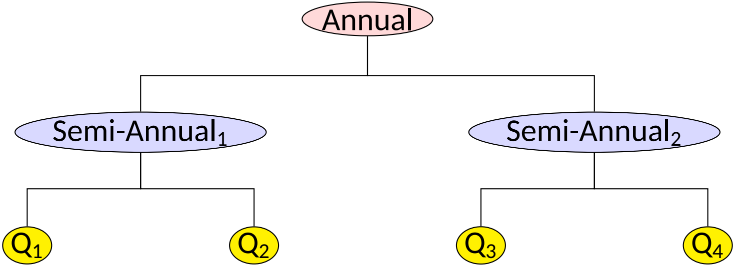
Temporal hierarchies



Basic idea:

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

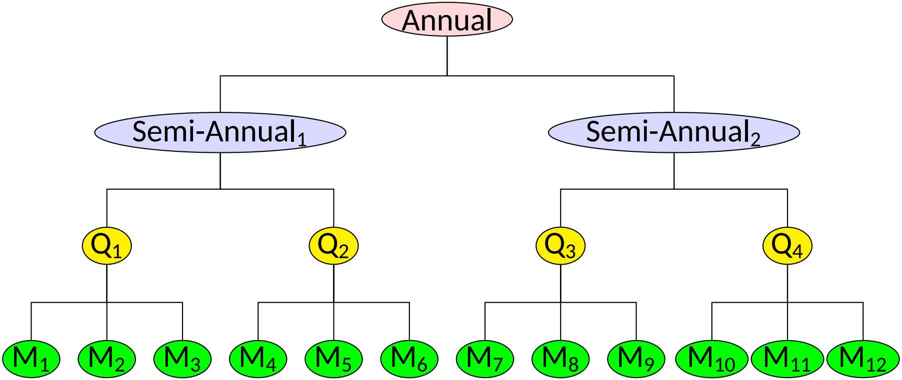
Temporal hierarchies



Basic idea:

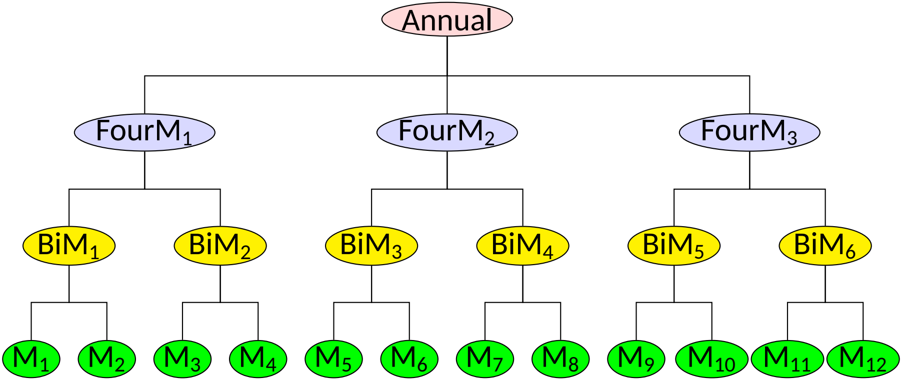
- Forecast series at each available frequency.
- Optimally reconcile forecasts within the same year.

Monthly series



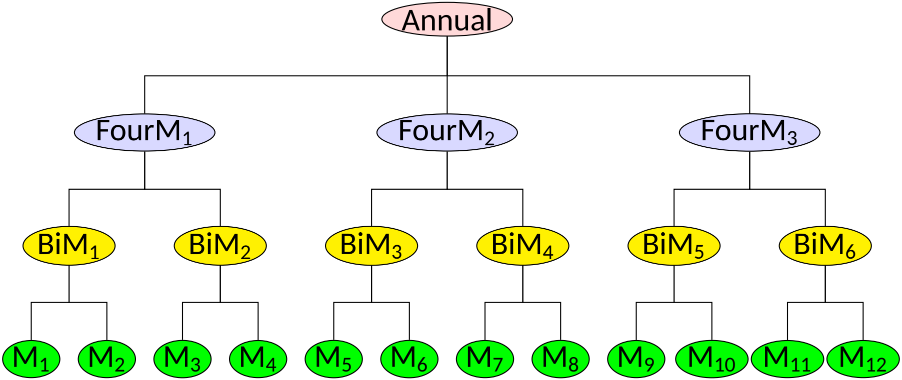
- $k = 2, 4, 12$ nodes
- $k = 3, 6, 12$ nodes
- Why not $k = 2, 3, 4, 6, 12$ nodes?

Monthly series



- $k = 2, 4, 12$ nodes
- $k = 3, 6, 12$ nodes
- Why not $k = 2, 3, 4, 6, 12$ nodes?

Monthly series



- $k = 2, 4, 12$ nodes
- $k = 3, 6, 12$ nodes
- Why not $k = 2, 3, 4, 6, 12$ nodes?

Monthly data

$$
\underbrace{\begin{pmatrix} A \\ SemiA_1 \\ SemiA_2 \\ FourM_1 \\ FourM_2 \\ FourM_3 \\ Q_1 \\ \vdots \\ Q_4 \\ BiM_1 \\ \vdots \\ BiM_6 \\ M_1 \\ \vdots \\ M_{12} \end{pmatrix}}_{(28\times 1)}
=
\underbrace{\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & \vdots & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & \vdots & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
 & & & & & & \boldsymbol{I}_{12} & & & & &
\end{pmatrix}}_{\boldsymbol{S}}
\underbrace{\begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \\ M_9 \\ M_{10} \\ M_{11} \\ M_{12} \end{pmatrix}}_{\boldsymbol{B}_t}
$$

In general

For a time series $y_1, \dots, y_T$, observed at frequency m, we generate aggregate series

$$y_j^{[k]} = \sum_{t=1+(j-1)k}^{jk} y_t, \qquad \text{for } j = 1, \dots, \lfloor T/k \rfloor$$

- $k \in F(m) = \{\text{factors of } m\}$.
- A single unique hierarchy is only possible when there are no coprime pairs in $F(m)$.
- $M_k = m/k$ is seasonal period of aggregated series.

In general

For a time series $y_1, \dots, y_T$, observed at frequency m, we generate aggregate series

$$y_j^{[k]} = \sum_{t=1+(j-1)k}^{jk} y_t, \qquad \text{for } j = 1, \dots, \lfloor T/k \rfloor$$

- $k \in F(m) = \{\text{factors of } m\}$.
- A single unique hierarchy is only possible when there are no coprime pairs in $F(m)$.
- $M_k = m/k$ is seasonal period of aggregated series.

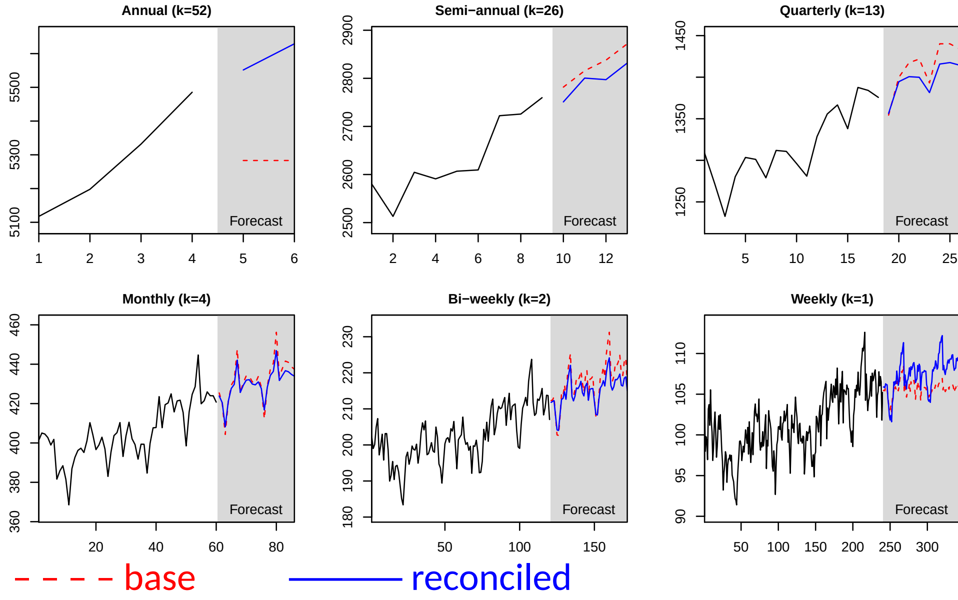
In general

For a time series $y_1, \dots, y_T$, observed at frequency m, we generate aggregate series

$$y_j^{[k]} = \sum_{t=1+(j-1)k}^{jk} y_t, \qquad \text{for } j = 1, \dots, \lfloor T/k \rfloor$$

- $k \in F(m) = \{\text{factors of } m\}$.
- A single unique hierarchy is only possible when there are no coprime pairs in $F(m)$.
- $M_k = m/k$ is seasonal period of aggregated series.

UK Accidents and Emergency Demand



UK Accidents and Emergency Demand

1. Type 1 Departments — Major A&E
2. Type 2 Departments — Single Specialty
3. Type 3 Departments — Other A&E/Minor Injury
4. Total Attendances
5. Type 1 Departments — Major A&E $>$ 4 hrs
6. Type 2 Departments — Single Specialty $>$ 4 hrs
7. Type 3 Departments — Other A&E/Minor Injury $>$ 4 hrs
8. Total Attendances $>$ 4 hrs
9. Emergency Admissions via Type 1 A&E
10. Total Emergency Admissions via A&E
11. Other Emergency Admissions (i.e., not via A&E)
12. Total Emergency Admissions
13. Number of patients spending $>$ 4 hrs from decision to admission

UK Accidents and Emergency Demand

- **Minimum training set**: all data except the last year
- Base forecasts using `auto.arima()`.
- Mean Absolute Scaled Errors for 1, 4 and 13 weeks ahead using a rolling origin.

Aggr. Level	h	Base	Reconciled	Change
Weekly	1	1.6	1.3	−17.2%
Weekly	4	1.9	1.5	−18.6%
Weekly	13	2.3	1.9	−16.2%
Weekly	1–52	2.0	1.9	−5.0%
Annual	1	3.4	1.9	−42.9%

UK Accidents and Emergency Demand

- **Minimum training set**: all data except the last year
- Base forecasts using `auto.arima()`.
- Mean Absolute Scaled Errors for 1, 4 and 13 weeks ahead using a rolling origin.

Aggr. Level	h	Base	Reconciled	Change
Weekly	1	1.6	1.3	−17.2%
Weekly	4	1.9	1.5	−18.6%
Weekly	13	2.3	1.9	−16.2%
Weekly	1–52	2.0	1.9	−5.0%
Annual	1	3.4	1.9	−42.9%

UK Accidents and Emergency Demand

- **Minimum training set**: all data except the last year
- Base forecasts using `auto.arima()`.
- Mean Absolute Scaled Errors for 1, 4 and 13 weeks ahead using a rolling origin.

Aggr. Level	h	Base	Reconciled	Change
Weekly	1	1.6	1.3	−17.2%
Weekly	4	1.9	1.5	−18.6%
Weekly	13	2.3	1.9	−16.2%
Weekly	1–52	2.0	1.9	−5.0%
Annual	1	3.4	1.9	−42.9%

UK Accidents and Emergency Demand

- **Minimum training set**: all data except the last year
- Base forecasts using `auto.arima()`.
- Mean Absolute Scaled Errors for 1, 4 and 13 weeks ahead using a rolling origin.

Aggr. Level	h	Base	Reconciled	Change
Weekly	1	1.6	1.3	−17.2%
Weekly	4	1.9	1.5	−18.6%
Weekly	13	2.3	1.9	−16.2%
Weekly	1–52	2.0	1.9	−5.0%
Annual	1	3.4	1.9	−42.9%

UK Accidents and Emergency Demand

- **Minimum training set**: all data except the last year
- Base forecasts using `auto.arima()`.
- Mean Absolute Scaled Errors for 1, 4 and 13 weeks ahead using a rolling origin.

Aggr. Level	h	**Base**	**Reconciled**	**Change**
Weekly	1	1.6	1.3	−17.2%
Weekly	4	1.9	1.5	−18.6%
Weekly	13	2.3	1.9	−16.2%
Weekly	1–52	2.0	1.9	−5.0%
Annual	1	3.4	1.9	−42.9%

thief package for R

Temporal Hierarchical Forecasting

Install from CRAN

```
install.packages("thief")
```

Usage

```
library(thief)
thief(y)
```

thief package for R

Temporal Hierarchical Forecasting

Install from CRAN

```
install.packages("thief")
```

Usage

```
library(thief)
thief(y)
```

Outline

Coherent density forecasts

Definition: Coherence

Suppose $\boldsymbol{y}_t \in \mathbb{R}^n$. $\boldsymbol{y}_t$ is *coherent* if $\boldsymbol{y}_t$ lies in an m-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of the summing matrix $\boldsymbol{S}$.

Definition: Coherent density forecasts

Any density $p(\boldsymbol{y}_{t+h})$ is coherent if $p(\boldsymbol{y}_{t+h}) = 0$ for all $\boldsymbol{y}_{t+h}$ in the null space of $\boldsymbol{S}$.

- Corollary: The probability distribution at each node is a convolution of the child distributions.
- Coherent point forecasts: $\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h}$.
- Coherent variance forecasts: $\tilde{\Sigma}_{T+h|T} = \boldsymbol{S}\boldsymbol{P}\hat{\Sigma}_{T+h|T}\boldsymbol{P}'\boldsymbol{S}'$.

Coherent density forecasts

Definition: Coherence

Suppose $\boldsymbol{y}_t \in \mathbb{R}^n$. $\boldsymbol{y}_t$ is *coherent* if $\boldsymbol{y}_t$ lies in an m-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of the summing matrix $\boldsymbol{S}$.

Definition: Coherent density forecasts

Any density $p(\boldsymbol{y}_{t+h})$ is coherent if $p(\boldsymbol{y}_{t+h}) = 0$ for all $\boldsymbol{y}_{t+h}$ in the null space of $\boldsymbol{S}$.

- Corollary: The probability distribution at each node is a convolution of the child distributions.
- Coherent point forecasts: $\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{SP}\hat{\boldsymbol{y}}_{T+h}$.
- Coherent variance forecasts: $\tilde{\Sigma}_{T+h|T} = \boldsymbol{SP}\hat{\Sigma}_{T+h|T}\boldsymbol{P}'\boldsymbol{S}'$.

Coherent density forecasts

Definition: Coherence

Suppose $\boldsymbol{y}_t \in \mathbb{R}^n$. $\boldsymbol{y}_t$ is *coherent* if $\boldsymbol{y}_t$ lies in an m-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of the summing matrix $\boldsymbol{S}$.

Definition: Coherent density forecasts

Any density $p(\boldsymbol{y}_{t+h})$ is coherent if $p(\boldsymbol{y}_{t+h}) = 0$ for all $\boldsymbol{y}_{t+h}$ in the null space of $\boldsymbol{S}$.

- Corollary: The probability distribution at each node is a convolution of the child distributions.
- Coherent point forecasts: $\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h}$.
- Coherent variance forecasts: $\tilde{\Sigma}_{T+h|T} = \boldsymbol{S}\boldsymbol{P}\hat{\Sigma}_{T+h|T}\boldsymbol{P}'\boldsymbol{S}'$.

Outline

Coherent Gaussian forecasts

$$\boldsymbol{y}_{T+h|T} \sim N(\tilde{\boldsymbol{y}}_{T+h|T}, \tilde{\Sigma}_{T+h|T})$$

Let L be the Energy Score (a proper scoring rule):

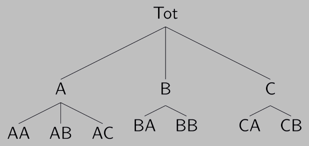
$$L(\tilde{F}_{T+h|T}, \boldsymbol{y}_{T+h}) = \mathrm{E}||\tilde{\boldsymbol{Y}}_{T+h} - \boldsymbol{y}_{T+h}||^{\alpha} - \frac{1}{2}\mathrm{E}||\tilde{\boldsymbol{Y}}_{T+h} - \tilde{\boldsymbol{Y}}'_{T+h}||^{\alpha}$$

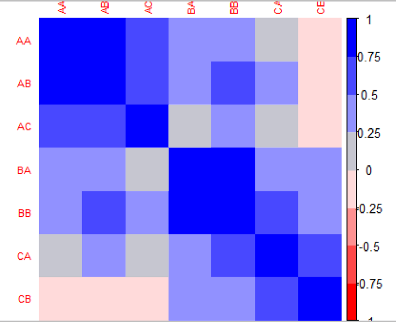
for $\alpha \in (0, 2]$, where $\tilde{\boldsymbol{Y}}_{T+h}$ and $\tilde{\boldsymbol{Y}}'_{T+h}$ are independent rvs from $\tilde{F}_{T+h|T} = N(\tilde{\boldsymbol{y}}_{T+h|T}, \tilde{\Sigma}_{T+h|T})$.

- There is no closed form expression for $L(\tilde{F}_{T+h|T}, \boldsymbol{y}_{T+h})$ for $\alpha \in (0, 2)$ under the Gaussian predictive distribution.
- When $\alpha = 2$, $L(\tilde{F}_{T+h|T}, \boldsymbol{y}_{T+h}) = \mathrm{E}||\tilde{\boldsymbol{y}}_{T+h|T} - \boldsymbol{y}_{T+h}||^2$
- This is equivalent to MinT solution.

Monte-Carlo simulation

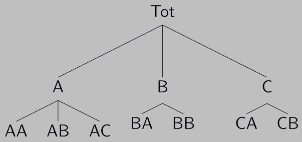
Hierarchy 1: Case A

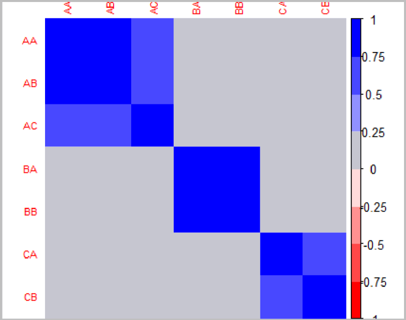




Monte-Carlo simulation

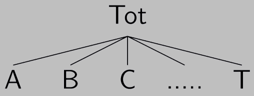
Hierarchy 1: Case B

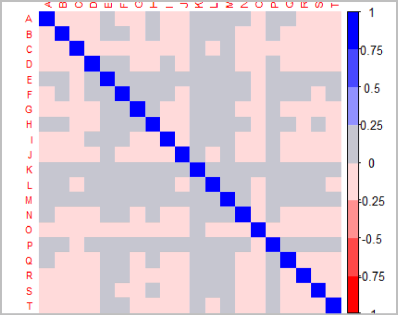




Monte-Carlo simulation

Hierarchy 2





Monte-Carlo simulation

- Bottom level series generated from univariate ARMA(1,1) processes.
- Contemporaneous errors randomly generated from multivariate Gaussian distribution with mean zero and correlation structures described before.
- Parameters for AR and MA components from uniform distribution, satisfying stationarity and invertibility conditions.

	Interval
Hierarchy 1: Case A	[0.4, 0.7]
Hierarchy 1: Case B	[0.4, 0.7]
Hierarchy 2	[0.3, 0.7]

Monte-Carlo simulation

- 501 observations generated for each series.
- Univariate ARIMA models fitted for first 500 observations and 1- step ahead base forecasts generated.
- Predictive means and variances obtained using different reconciliation methods.
- Process replicated 1000 times from same DGP.

Monte-Carlo simulation

- 501 observations generated for each series.
- Univariate ARIMA models fitted for first 500 observations and 1- step ahead base forecasts generated.
- Predictive means and variances obtained using different reconciliation methods.
- Process replicated 1000 times from same DGP.

	Average Energy Score		
Reconciliation method	Hierarchy 1A	Hierarchy 1B	Hierarchy 2
Base	9.26	6.65	9.76
Bottom up	9.19**	6.63	9.57**
OLS	9.23**	6.63**	9.74**
MinT(Sample)	9.20*	6.66	9.58**
MinT(Shrink)	9.19**	6.62**	9.60**

Monte-Carlo simulation

- 501 observations generated for each series.
- Univariate ARIMA models fitted for first 500 observations and 1- step ahead base forecasts generated.
- Predictive means and variances obtained using different reconciliation methods.
- Process replicated 1000 times from same DGP.

	Diebold-Mariano test: best pairwise method		
Reconciliation method	Hierarchy 1A	Hierarchy 1B	Hierarchy 2
BU vs OLS			BU
BU vs MinT(Sample)		BU	
BU vs MinT(Shrink)			
OLS vs MinT(Sample)			MinT(Sample)
OLS vs MinT(Shrink)	MinT(Shrink)		MinT(Shrink)
MinT(Shrink) vs MinT(Sample)		MinT(Shrink)	

Outline

Coherent nonparametric forecasts

1. Fit univariate models at each node using data up to time T.
2. Let $\boldsymbol{R} = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_T)'$ be a matrix of residuals where $\boldsymbol{e}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$.
3. Let $\boldsymbol{E}^b = (\boldsymbol{e}_{i+1}, \dots, \boldsymbol{e}_{i+h})'$ be a block bootstrap sample of size h from $\boldsymbol{R}$.
4. Generate h-step ahead sample paths from the fitted models incorporating $\boldsymbol{E}^b$. Denote by $\boldsymbol{y}^b_{T+h}$.
5. Project sample paths to coherent space: $\tilde{\boldsymbol{y}}^b_{T+h} = \boldsymbol{SPy}^b_{T+h}$ where $\tilde{\boldsymbol{y}}^b_{T+h}$ denote coherent h-step ahead sample paths.
6. Repeat step 3–5 J times.

Monte-Carlo simulation

- 501 observations generated for each series.
- Univariate ARIMA models fitted for first 500 observations and 1- step ahead base forecasts generated.
- 5000 1-step future paths constructed for 500 replications from same DGP.

Monte-Carlo simulation

- 501 observations generated for each series.
- Univariate ARIMA models fitted for first 500 observations and 1- step ahead base forecasts generated.
- 5000 1-step future paths constructed for 500 replications from same DGP.

	Average Energy Score		
Reconciliation method	Hierarchy 1A	Hierarchy 1B	Hierarchy 2
Base	14.54	12.44	13.59
Bottom up	13.87**	11.76**	13.77
OLS	14.17**	12.11**	13.53**
MinT(Sample)	15.12	12.98	13.61
MinT(Shrink)	14.15**	12.15**	13.37**

Monte-Carlo simulation

- 501 observations generated for each series.
- Univariate ARIMA models fitted for first 500 observations and 1- step ahead base forecasts generated.
- 5000 1-step future paths constructed for 500 replications from same DGP.

	Diebold-Mariano test: best pairwise method		
Reconciliation method	Hierarchy 1A	Hierarchy 1B	Hierarchy 2
BU vs OLS	BU	BU	
BU vs MinT(Sample)	BU	BU	MinT(Sample)
BU vs MinT(Shrink)	BU	BU	MinT(Shrink)
OLS vs MinT(Sample)	OLS	OLS	
OLS vs MinT(Shrink)			MinT(Shrink)
MinT(Shrink) vs MinT(Sample)	MinT(Shrink)	MinT(Shrink)	MinT(Shrink)

Copula-based distributions of sums

Sklar's theorem

For any continuous distribution $\boldsymbol{F}$ with marginals $F_1, \ldots, F_d$, there exists a unique "copula" function $\mathbf{C} : [0,1]^d \to [0,1]$ such that

$$\boldsymbol{F}(x_1, \ldots, x_d) = \mathbf{C}(F_1(x_1), \ldots, F_d(x_d))$$

Empirical copula

If $x_k^i \sim F_i$ and $\boldsymbol{u}_k = (u_k^1, \ldots, u_k^d) \sim \mathbf{C}$, then

$$\hat{F}_i(x) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{x_k^i \leq x\}$$

and empirical copula is

$$\mathbf{C}(\boldsymbol{u}) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\left\{\frac{rk(u_k^1)}{K} \leq u_1, \ldots, \frac{rk(u_k^d)}{K} \leq u_d\right\}$$

$$\hat{F}(x_1, \ldots, x_d) = \hat{C}(\hat{F}_1(x_1), dots, \hat{F}_d(x_d))$$

Copula-based distributions of sums

- We can efficiently compute $\hat{F}$ using permutations.
- We can compute copulas recursively in the tree structure, rather than find the joint distribution or the entire hierarchy.

Copula-based distributions of sums

- We can efficiently compute $\hat{\boldsymbol{F}}$ using permutations.
- We can compute copulas recursively in the tree structure, rather than find the joint distribution or the entire hierarchy.

Coherent nonparametric forecasts

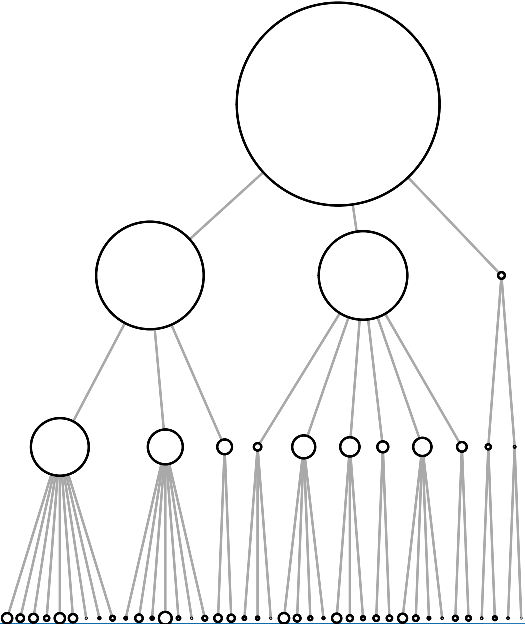
1. Forecast at every node using whatever method you choose to get marginal forecast distributions for each node.
2. Apply MinT to reconcile the means of the forecast distributions.
3. Simulate from the forecast distributions at each bottom level node.
4. Compute empirical copulas for each parent+children group to obtain coherent forecast distributions at the next level up.
5. Repeat working up the tree.

Application: Smart Meter Data

Figure: `http://solutions.3m.com`

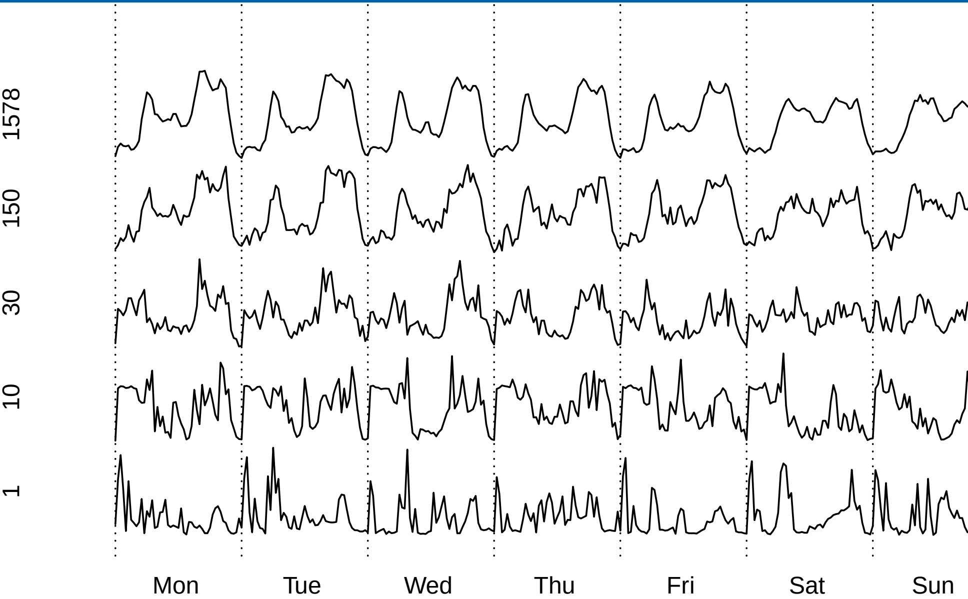
- 1578 households from Great Britain.
- Half-hourly data from 20 April 2009 – 31 July 2010.
- Training data: to 30 April 2010.
- Forecasting 48 periods ahead (one day).
- Geographical hierarchy with five levels.

Application: Smart Meter Data

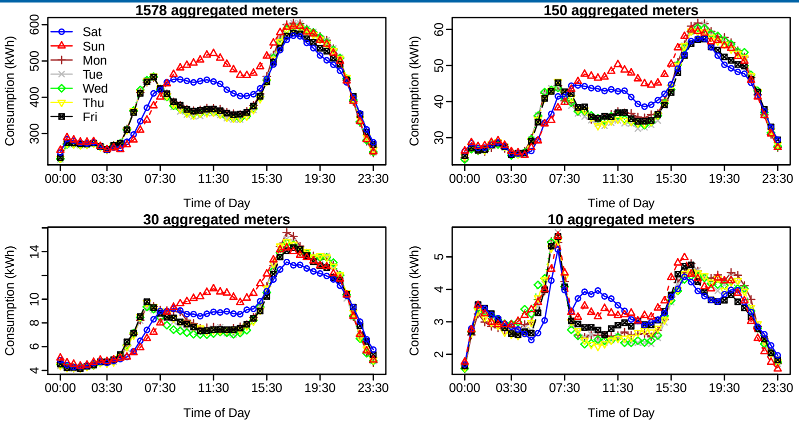


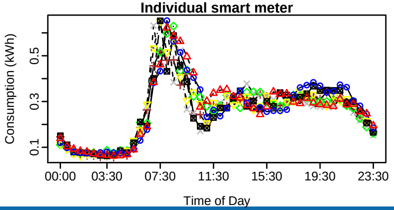
- 3 groups at level 2
- 11 groups at level 3
- 40 groups at level 4.
- 1578 households at bottom level.

Application: Smart Meter Data



Application: Smart Meter Data



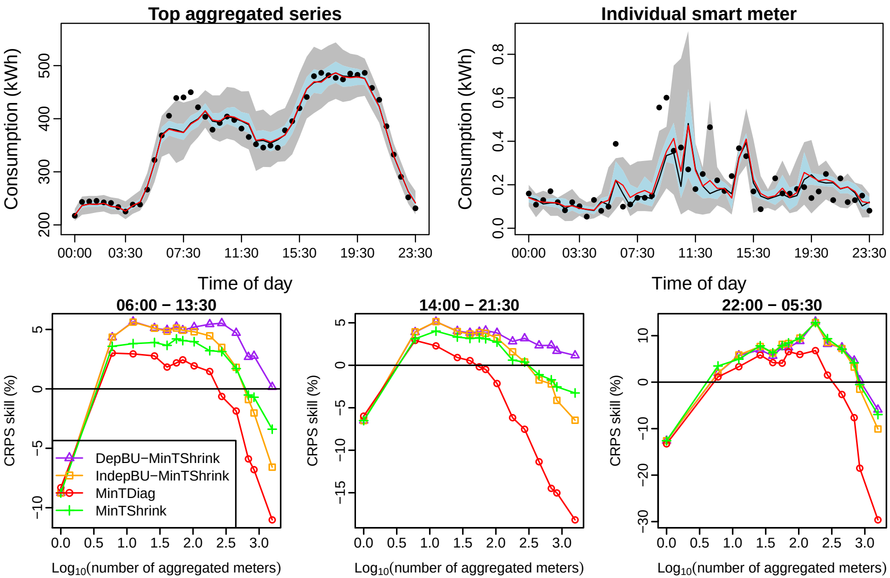


One week of demand at different levels of aggregation.

Application: Smart Meter Data

- Forecast individual series using Taylor's double-seasonal Holt-Winters' method.
- Kernel density estimation by 48 half-hours and for 3 different day types (weekday, Saturday, Sunday) for density forecasts.
- KDE use decay parameter to "forget" the past.
- Decay and bandwidth chosen to minimize CRPS

Application: Smart Meter Data



CRPS skill relative to base forecasts.

Outline

Conclusions

- MinT (Shink) not only optimally reconciles point forecasts, it is also optimal for probabilistic Gaussian forecasts.
- MinT (Shrink) can also be used to generate coherent future sample paths.
- Combining MinT (Shrink) with empirical copulas allows for efficient nonparametric coherent probabilistic forecasting.

References

G Athanasopoulos, RJ Hyndman, N Kourentzes and F Petropoulos (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research* **262**(1), 60–74

SL Wickramasuriya, G Athanasopoulos and RJ Hyndman (2015). *Forecasting hierarchical and grouped time series through trace minimization*. Working paper. Dept Econometrics & Business Statistics, Monash University

S Ben Taieb, JW Taylor and RJ Hyndman (2017). *Hierarchical Probabilistic Forecasting of Electricity Demand with Smart Meter Data*. Working paper. Dept Econometrics & Business Statistics, Monash University

Plus contributions from: Anastasios Panagiotelis, George Athanasopoulos, Puwasala Gamakumara.