COMPSTAT 2010 19° International Conference
on Computational Statistics
Paris-France, August 22-27

Forecasting Complex Time Series:
Beanplot Time Series

Carlo Drago and Germana Scepi

Dipartimento di Matematica e Statistica
Università "Federico II" di Napoli

The Aim

☞Dealing with "complex" time series:

Visualizing *(CLADAG 2009,Gfkl 2010)*

Scalar Time Series → Bean Plot Time Series

Synthesizing the global dynamics

Beanplot Time Series → Parametrization → AttributeTime Series

Forecasting beanplot dynamics

Attribute Time Series → Forecasting Beanplot dynamics

Complex time series

☞ **"complex" time series: Financial Time Series**

Higher Volatility

Structural Changes

Volatility Clustering

High Frequency data: **the number of observations can be overwhelming with periodic (intra-day and intra-week) patterns**

Irregularly spaced time series with random daily numbers of observations

Missing data

Visualizing, modeling and forecasting

Beanplot time series

☞ **A beanplot time series is an ordered sequence of beanplots over the time. Each temporal interval can be considered as a domain of values that is related to the chosen interval temporal (daily, week, and month).**

The beanplot can be considered as a particular case of an interval-valued modal variable at the same time like boxplots and histograms (see *Arroyo and Mate 2006)*

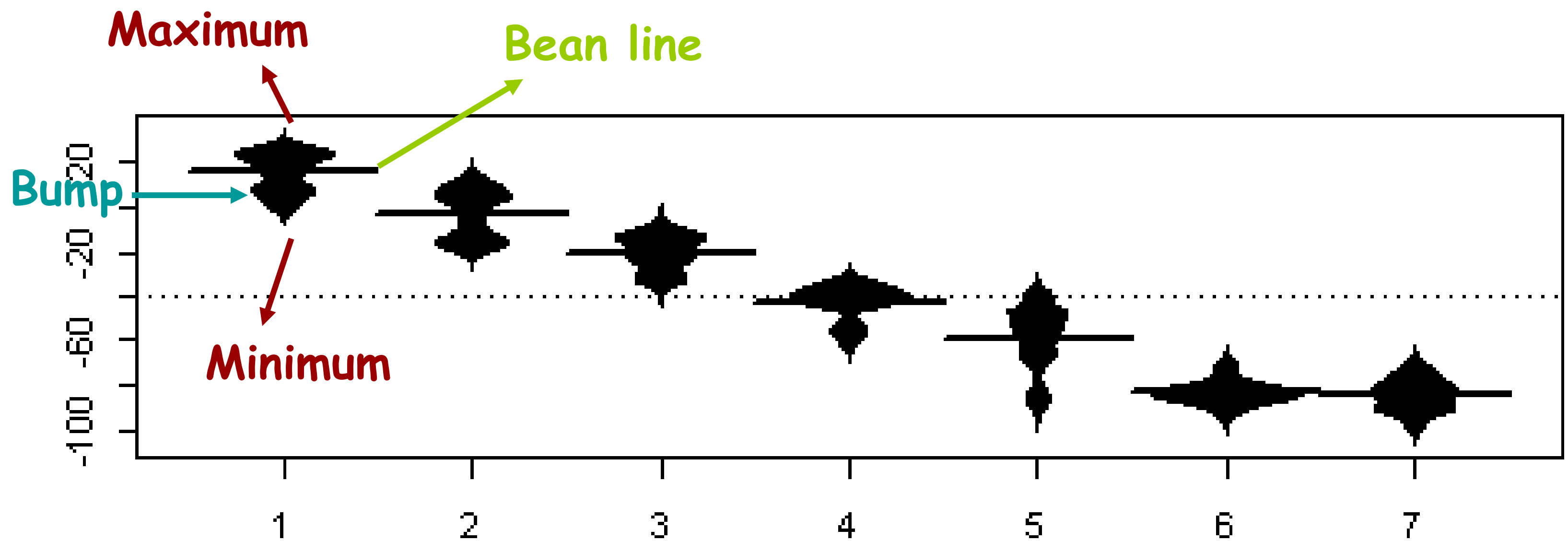
In a beanplot variable we are taking into account at the same time the intervals of minimum and maximum and the density in form of a kernel nonparametric estimator (the density trace see *Kampstra 2008*).

Kernel

$$\hat{f}_{x,h} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

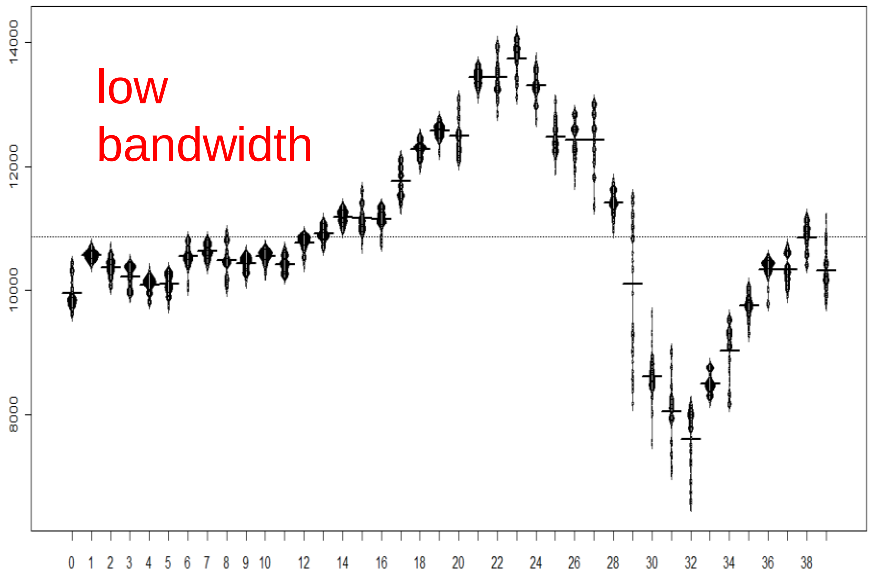
Bandwidth

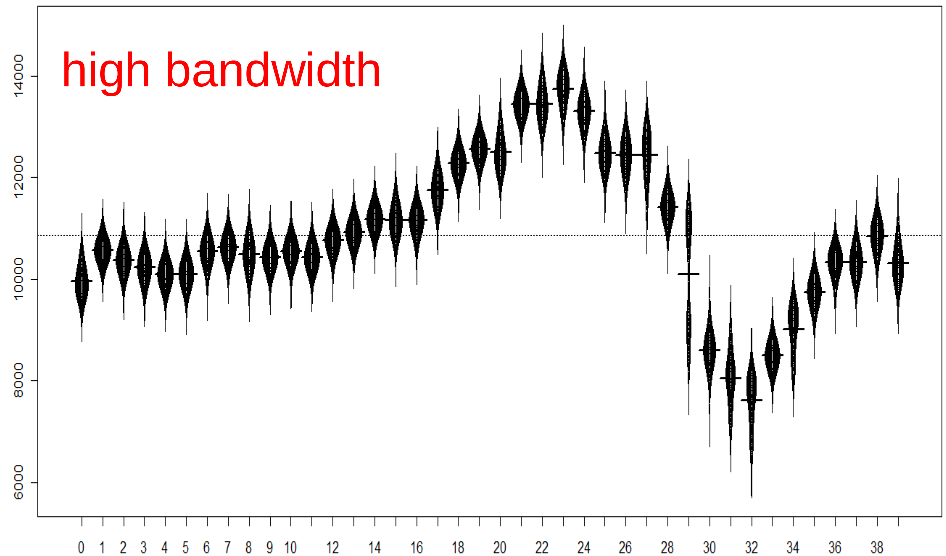
Beanplot time series

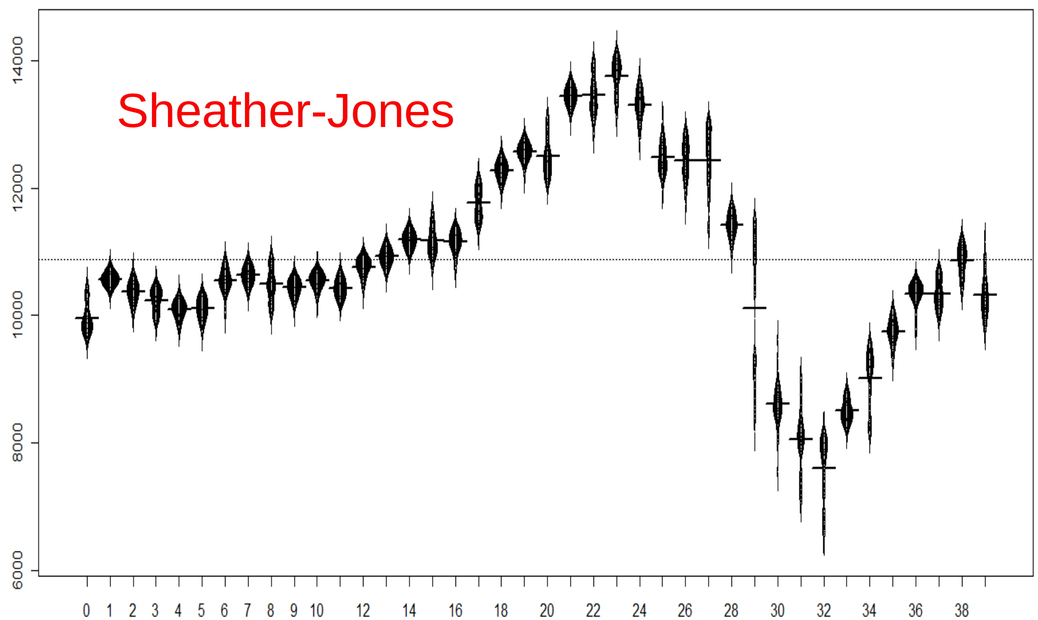


The beanplot time series show the complex structure of the underlying phenomenon by representing jointly the data *location (the bean line)* the *size* (*the interval between minimum and maximum*) and the *shape* (the *density trace*) over the time

The **bumps** represent the values of maximum density showing important equilibrium values reached in a single temporal interval. Bumps can also show the intra-period patterns over the time and more in general the beanplot *shape* shows the *intra-period* dynamic







Dow Jones closing prices from the 1-11-2003 to the 30-6-2010

Attribute time series

☞ **For each time t we consider** ***an internal model represented by each Beanplot***

☞ **For each time t we can consider *n* descriptors of the beanplots**

☞ **Each descriptor is represented over the time as an attribute time series (see *Matè and Arroyo ,2008*)**

☞ **By the attribute time series we take into account the dynamics of the phenomenon. In this sense we can consider the correlation over the time of the beanplot features**

Attribute time series (1)

☞ **At each time t from the kernel density estimate we consider the minimum, maximum, center and some coefficients from a polynomial model.**

Data: A Beanplot time series $\{b_{Y_t}\}\ t = 1...T$
Result: A vector with n elements of $\bar{x}$ (minima, maxima and a center) and $\bar{y}$ (the polynomial coefficients) given the bandwidth h

begin
  Choice of the n points to parametrize
  Choice of the h bandwidth to use
  for $t \in T$ **do**
    Parametrizing the $\bar{x}$ by extracting the minima, maxima and centers
    Parametrizing the $\bar{y}$ by estimating a polynomial regression
  end
  Is the internal models not fitting data adequately?
  if *the internal model is not adequated fitted* **then**
    change the number of parameters n or the bandwidth h
  end
end

Algorithm 1: Beanplot internal modelling: parametrization

Attribute time series (1)

☞ At each time t from the kernel density estimate we consider the minimum, maximum, center and some coefficients from a polynomial model.

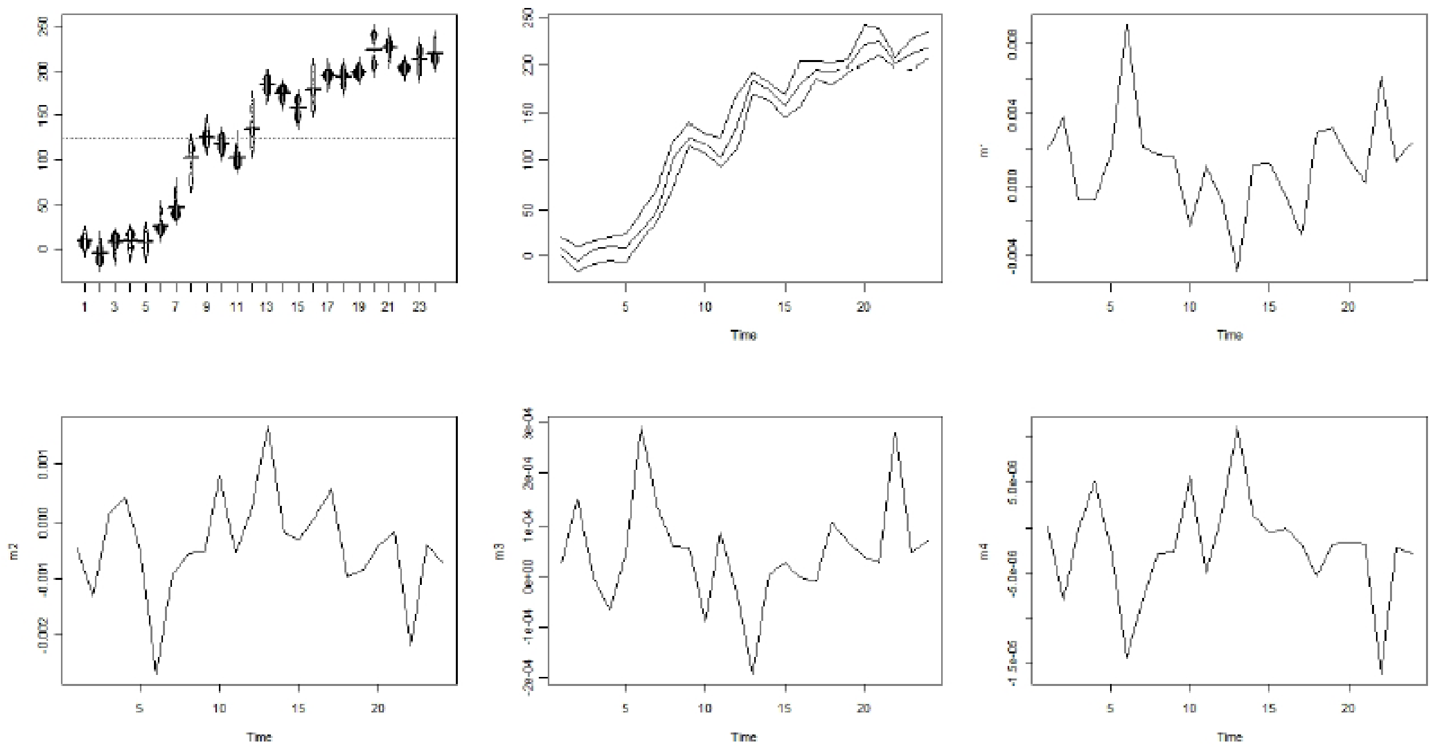


Fig. 4. A simulated beanplot time series: time parametrizations.

Attribute time series (2)

- **Alternative: at each time t from the kernel density estimate we can obtain *n* parameters as coordinates** $\bar{x}$ $\bar{y}$

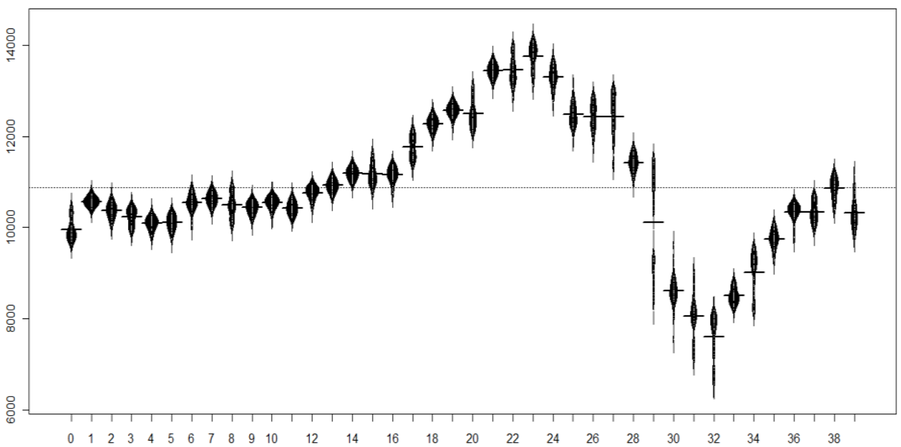
Data: A Beanplot time series $\{b_{Y_t}\}\ t = 1...T$
Result: A vector with n elements of $\bar{x}$ and $\bar{y}$ coordinates given the bandwidth h

begin
  Choice of the n points to parametrize
  Choice of the h bandwidth to use
  for $t \in T$ **do**
    Parametrizing the $\bar{x}$
    Parametrizing the $\bar{y}$
  end
  Is the internal models not fitting data adequately?
  if *the internal model is not adequated fitted* **then**
    change the number of parameters n or the bandwidth h
  end
end

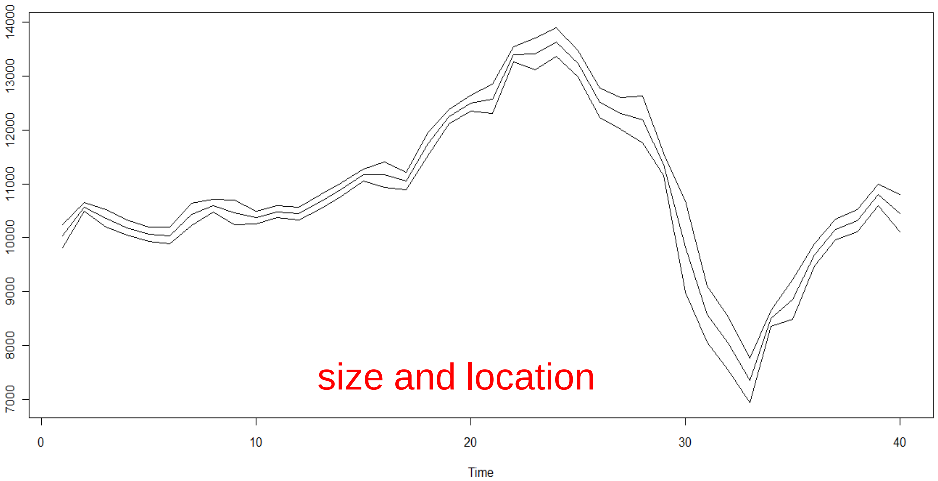
Algorithm 2: Beanplot internal modelling: parametrization

Parametrization example: Dow Jones data

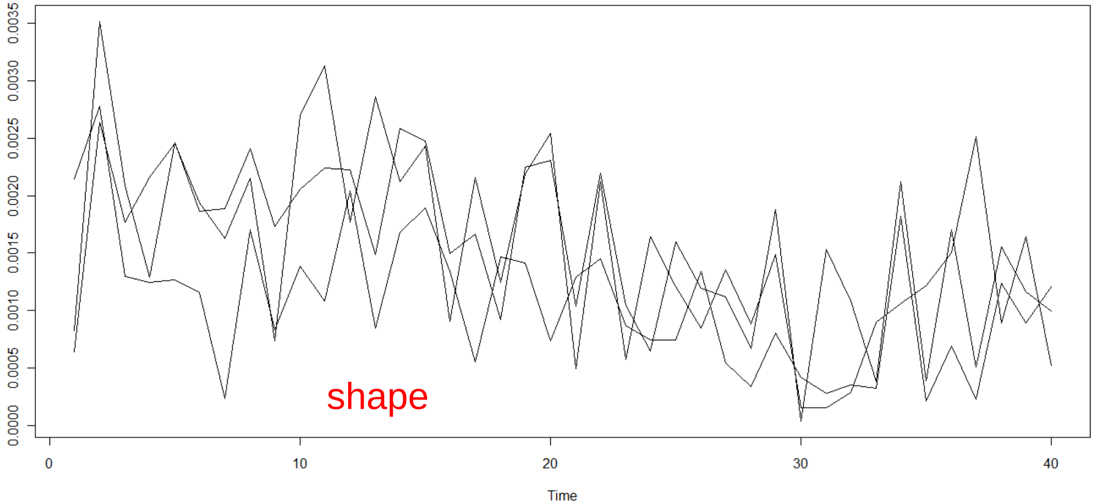
Beanplot time series for the closing prices



Attribute time series (X; 25; 50;75)



Attribute time series (Y;25;50;75)



Dow Jones closing prices from the 1-11-2003 to the 30-6-2010

The bandwidth chosen and used in the application is h=80.

External Models

☞ **Start to consider the *n* attribute time series of the descriptors (e.g. x1,x2,x3,y1,y2,y3) of the beanplots for t=1,...,T**

☞ **The attribute time series represent the external models (the dynamics over the time t=1,...,T) where each beanplot can be considered as the internal model at time t**

⇓

Forecasting

attribute time series

Forecasting methods

- ☞ **Univariate Methods (ARIMA, Smoothing Splines, Neural Networks, Hybrid Methods)**
- ☞ **Multivariate Methods (VAR, VECM)**
- ☞ **Forecasts combination**

Univariate methods when there is not an explicit relationship between the attributes with/or without autocorrelation

Multivariate methods if a correlation explicitly exists

Forecasting Procedure

- Start to consider the *n* attribute time series of the descriptors of the beanplots for t=1,...,T. They represent the beanplot dynamics over the time

- Checking for the stationarity and the autocorrelation. Detecting the features of the dynamics (trends, cycles, seasonality). Analyzing the relationships between the attributes

- Forecasting them using a specific method

- Considering as Beanplot description the forecasts obtained from the Forecasting Method.

- Diagnostics

Forecasting on attribute (coordinates) time series

- ☞ Start to consider the *n* attribute time series of coordinates

- ☞ Checking the autocorrelation in the X and in the Y. Analyzing the relationships between the X and between Y. Analyzing the features of the dynamics (trends, cycles, seasonality).

- ☞ Choose one or two methods of forecasting for X and Y.

- ☞ Considering as Beanplot description the forecasts obtained from the Forecasting Method.

- ☞ Diagnostics

We have tested our procedure on a lot of simulated data sets, with high number of observations and different starting models, we report only the results obtained on the real data set of Dow Jones

Application

- Dow Jones data (1928-10-01\2010-7-30 – 20549 observations)

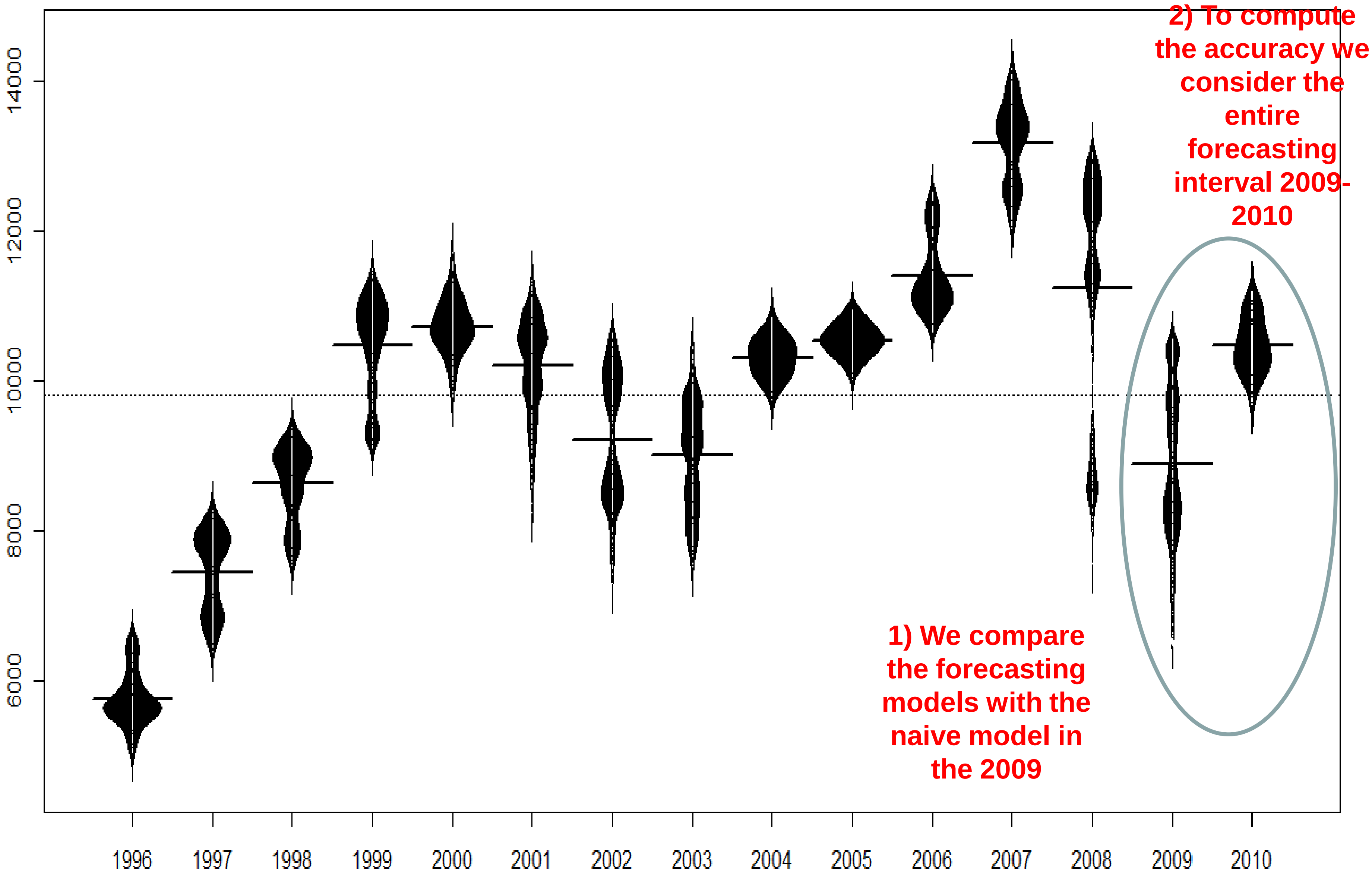
- Forecasting model period (1998-08-03\2008-08-03). Forecasting of the 2009 year and for the interval 2009-2010

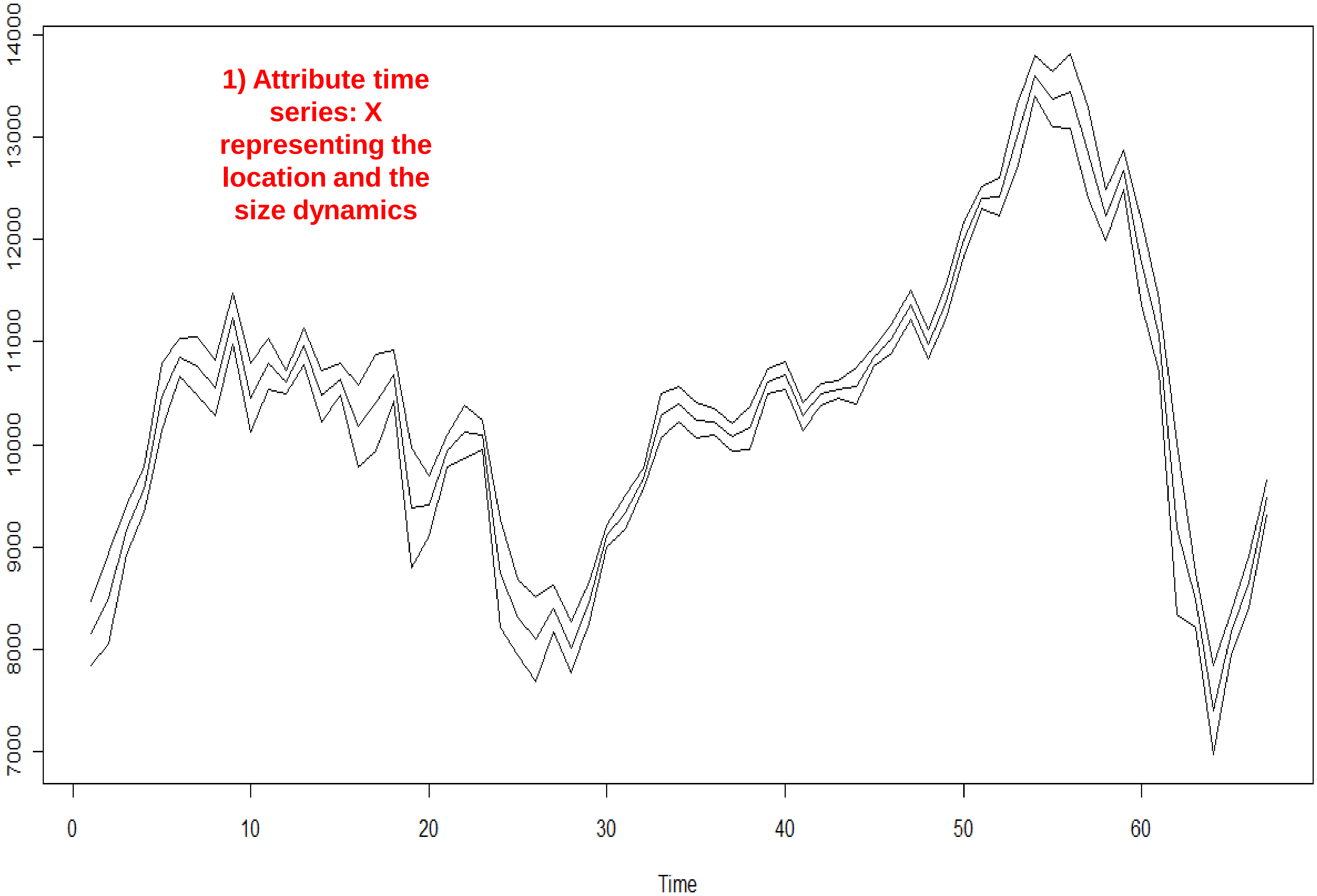
- Forecasting methods used: VAR, Auto-Arima, Exponential Smoothing, Smoothing Splines.

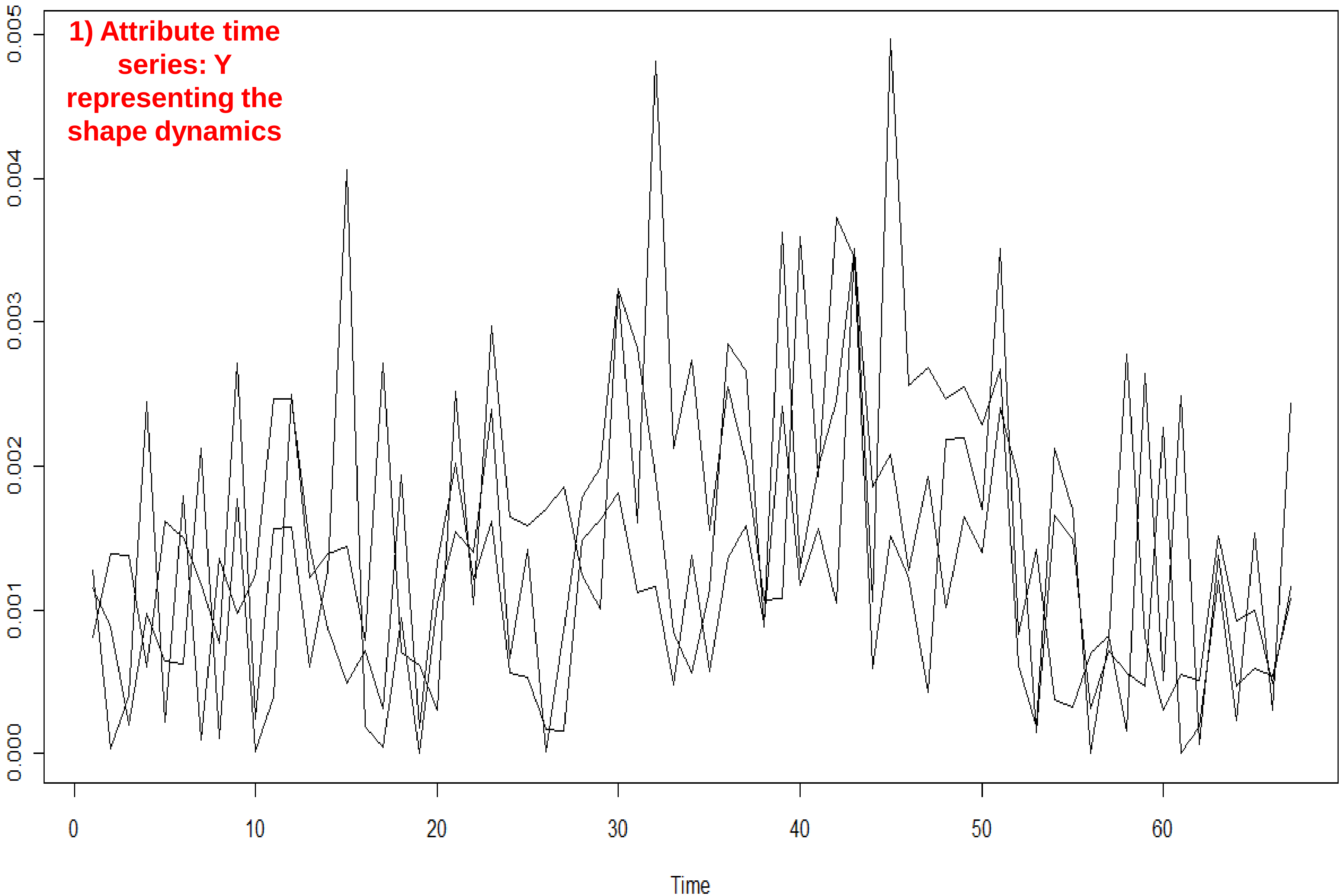
- Forecasting combinations (Mean, Exponential Smoothing, Auto-Arima) …

- Comparing the forecasts obtained with whose obtained by the “naïve” model

- Diagnostics (accuracy)


14000
12000
10000
8000
6000
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2) To compute the accuracy we consider the entire forecasting interval 2009-2010
1) We compare the forecasting models with the naive model in the 2009


1) Attribute time series: X representing the location and the size dynamics
14000
13000
12000
11000
10000
9000
8000
7000
0
10
20
30
40
50
60
Time


1) Attribute time series: Y representing the shape dynamics
0.000
0.001
0.002
0.003
0.004
0.005
0
10
20
30
40
50
60
Time

Augmented-Dickey-Fuller tests on the attribute time series (1)

1) X

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$x1[420:474])
Dickey-Fuller = -1.2098, Lag order = 3, p-value = 0.893
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$x2[420:474])
Dickey-Fuller = -1.0624, Lag order = 3, p-value = 0.9203
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$x3[420:474])
Dickey-Fuller = -0.9409, Lag order = 3, p-value = 0.9393
alternative hypothesis: stationary
```

1) Y

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$y1[420:474])
Dickey-Fuller = -4.1904, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(as.vector(d$y1[420:474])) :
  p-value smaller than printed p-value
```

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$y2[420:474])
Dickey-Fuller = -3.5076, Lag order = 3, p-value = 0.04903
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$y3[420:474])
Dickey-Fuller = -3.4256, Lag order = 3, p-value = 0.06106
alternative hypothesis: stationary
```

Augmented-Dickey-Fuller tests on the attribute time series (2)

1) X

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$x1[420:486])
Dickey-Fuller = -2.6031, Lag order = 4, p-value = 0.3303
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$x2[420:486])
Dickey-Fuller = -2.837, Lag order = 4, p-value = 0.2353
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$x3[420:486])
Dickey-Fuller = -3.0299, Lag order = 4, p-value = 0.1570
alternative hypothesis: stationary
```

1) Y

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$y1[420:486])
Dickey-Fuller = -3.1761, Lag order = 4, p-value = 0.09909
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$y2[420:486])
Dickey-Fuller = -2.6417, Lag order = 4, p-value = 0.3146
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  as.vector(d$y3[420:486])
Dickey-Fuller = -2.9976, Lag order = 4, p-value = 0.1701
alternative hypothesis: stationary
```

X- Attribute Time Series Phillips-Ouliaris Cointegration test

Year 1998-2008

```
        Phillips-Ouliaris Cointegration Test

data:  x
Phillips-Ouliaris demeaned = -66.6506, Truncation lag parameter = 0, p-value = 0.01

Warning message:
In po.test(x) : p-value smaller than printed p-value
```

All observations

```
        Phillips-Ouliaris Cointegration Test

data:  x
Phillips-Ouliaris demeaned = -476.8951, Truncation lag parameter = 4, p-value = 0.01

Warning message:
In po.test(x) : p-value smaller than printed p-value
```

X- Attribute Time Series Forecasting Model: Smoothing Splines

```
> fcast
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
68       10381.40 9714.439 11048.37 9361.369 11401.44
>
> ma[487:487]
[1] 10217.75
>
> ma[486:486]
[1] 9654.484
>
```

```
> fcast
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
68       10186.09 9434.691 10937.48 9036.927 11335.24
>
> me[487:487]
[1] 9971.102
>
> me[486:486]
[1] 9479.097
>
```

```
> fcast
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
68       9929.244 9051.217 10807.27 8586.417 11272.07
>
> mi[487:487]
[1] 9724.453
>
> mi[486:486]
[1] 9303.71
```

X- Attribute Time Series Forecasting Model: Auto-Arima

```
   Point Forecast    Lo 80   Hi 80    Lo 95    Hi 95
68        9943.349 9344.497 10542.2 9027.483 10859.21
>
> ma[487:487]
[1] 10217.75
>
> ma[486:486]
[1] 9654.484
>
```

```
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
68        9932.239 9282.696 10581.78 8938.849 10925.63
>
> me[487:487]
[1] 9971.102
>
> me[486:486]
[1] 9479.097
```

```
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
68        9303.71 8487.453 10119.97 8055.352 10552.07
>
> mi[487:487]
[1] 9724.453
>
> mi[486:486]
[1] 9303.71
```

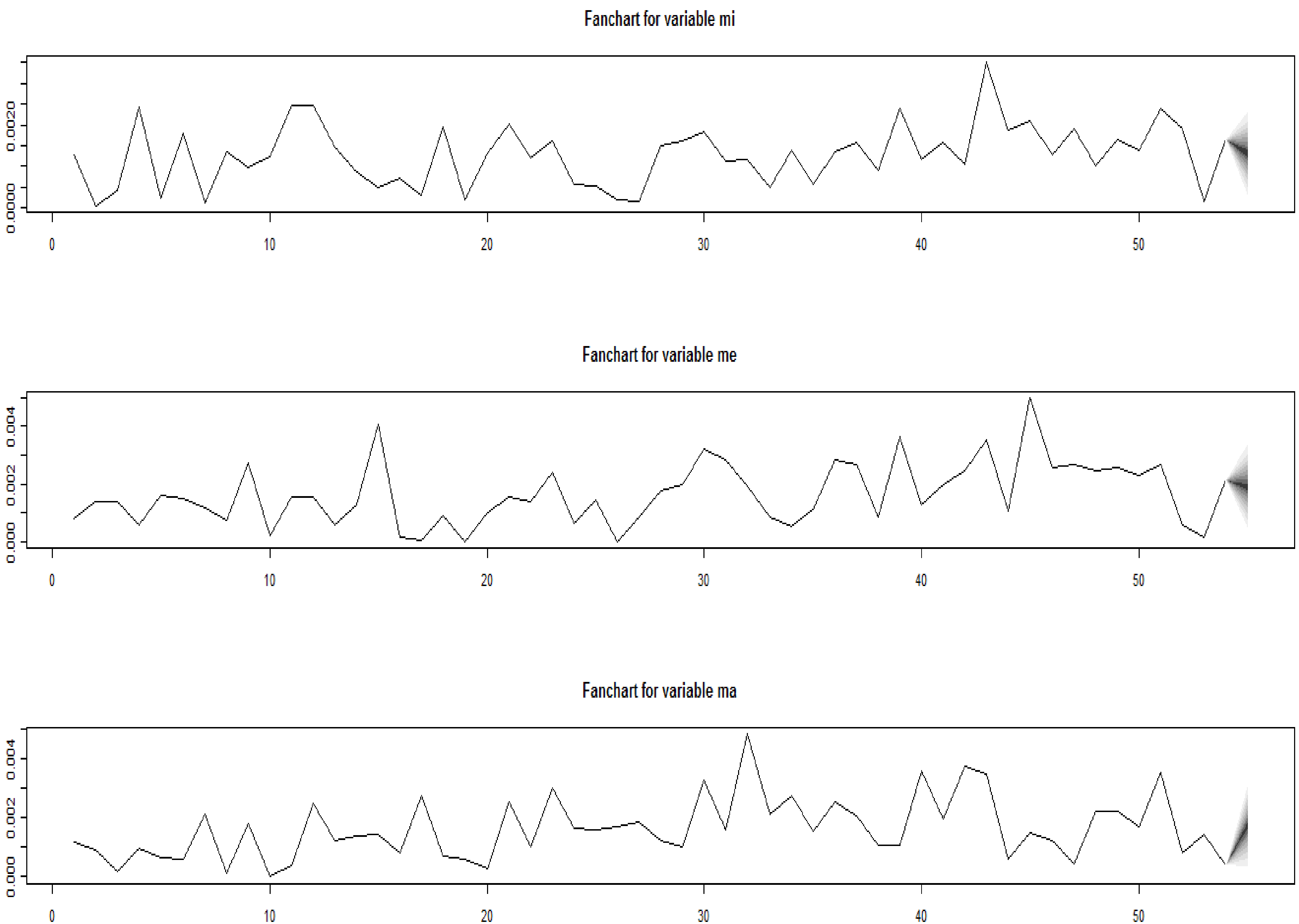
Y- Attribute Time Series Forecasting Model (1): VAR

```
$mi
               fcst          lower       upper          CI
mi.fcst 0.00130732 -0.0002138551 0.002828495 0.001521175

$me
                fcst         lower       upper          CI
me.fcst 0.001921221 -0.000262222 0.004104663 0.002183443

$ma
                fcst          lower       upper          CI
ma.fcst 0.001723613 -0.0003721843 0.003819411 0.002095797

>
> a[474:474,]
             mi          me            ma
474 0.001493228 0.001697756 0.0003257626
>
> a[473:473,]
             mi          me            ma
473 0.001662459 0.002125261 0.0003807355
```


Fanchart for variable mi
0.0000
0.0020
0
10
20
30
40
50
Fanchart for variable me
0.000
0.002
0.004
0
10
20
30
40
50
Fanchart for variable ma
0.000
0.002
0.004
0
10
20
30
40
50

Y- Attribute Time Series Forecasting Model (2): Smoothing Splines

```
> fcast
   Point Forecast         Lo 80       Hi 80        Lo 95       Hi 95
68   0.0006241531 -0.0006723961 0.001920702 -0.001358748 0.002607054
69   0.0005705273 -0.0007465263 0.001887581 -0.001443732 0.002584787
70   0.0005169014 -0.0008237607 0.001857563 -0.001533464 0.002567267
>
> ma[487:487]
[1] 0.0008703803
>
> ma[486:486]
[1] 0.001082339
```

```
> fcast
   Point Forecast         Lo 80       Hi 80        Lo 95       Hi 95
68   0.0005747982 -0.0007846799 0.001934276 -0.001504344 0.002653941
69   0.0005042496 -0.0008866913 0.001895191 -0.001623011 0.002631510
70   0.0004337011 -0.0009942359 0.001861638 -0.001750140 0.002617542
>
> me[487:487]
[1] 0.001508259
>
> me[486:486]
[1] 0.001163507
```

```
> fcast
   Point Forecast         Lo 80       Hi 80         Lo 95       Hi 95
68    0.001043531  6.708196e-06 0.002080354 -0.0005421527 0.002629215
69    0.001029806 -1.391073e-05 0.002073523 -0.0005664211 0.002626033
70    0.001016081 -3.525033e-05 0.002067413 -0.0005917916 0.002623954
>
> mi[487:487]
[1] 0.001308862
>
> mi[486:486]
[1] 0.002439116
```

Accuracy of the X - Forecasting Model: Smoothing Splines

Me

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-724.98359618	870.80098344	724.98359618	-7.00068426	7.00068426	1.66883151	-0.02703744

Theil's U
3.41615261

Mi

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-687.14544074	845.05524562	687.14544074	-6.78688151	6.78688151	1.44464157	-0.05126588

Theil's U
2.73143851

Ma

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-670.10896403	807.55987717	670.10896403	-6.32769632	6.32769632	1.63633285	-0.01099898

Theil's U
3.82159500

Accuracy of the X - Forecasting Model: Auto-Arima

Me

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
-26.7652053	108.0789029	89.5969885	-0.2517275	0.8668322	0.2062423	-0.2051609	0.4178992

Mi

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
730.2562626	762.9834507	730.2562626	7.2321626	7.2321626	1.5352769	-0.2538137	2.4132294

Ma

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
444.42729015	460.89262417	444.42729015	4.22206220	4.22206220	1.08524287	-0.07776208	2.08105251

Accuracy of the Y - Forecasting Model: VAR

```
$mi
            fcst         lower       upper          CI
[1,] 0.001307320 -0.0002138551 0.002828495 0.001521175
[2,] 0.001306081 -0.0002335895 0.002845752 0.001539671
[3,] 0.001281202 -0.0002590280 0.002821432 0.001540230

$me
            fcst         lower       upper          CI
[1,] 0.001921221 -0.0002622220 0.004104663 0.002183443
[2,] 0.001708049 -0.0005095786 0.003925676 0.002217627
[3,] 0.001680837 -0.0005372168 0.003898892 0.002218054

$ma
            fcst         lower       upper          CI
[1,] 0.001723613 -0.0003721843 0.003819411 0.002095797
[2,] 0.001677462 -0.0004598880 0.003814812 0.002137350
[3,] 0.001644710 -0.0004941241 0.003783543 0.002138834
```

```
> a[474:474,]
             mi          me           ma
474 0.001493228 0.001697756 0.0003257626
```

Forecasting Combinations

Mi

```
> fcastc
[1]  9663.423  9864.035 10064.648 10265.261 10465.873
>
> mi[487:487]
[1] 9724.453
>
> mi[486:486]
[1] 9303.71
>
> accuracy(fcastc, mi[487:491], test=1:length(mi[487:491]))
         ME        RMSE         MAE         MPE        MAPE        ACF1   Theil's U
-11.1463398 306.9813503 218.0693273  -0.1394663   2.1775206   0.1160837   1.0390254
```

Ma

```
> fcastc
[1] 10148.26 10370.51 10575.59 10773.98 10969.77
>
> ma[487:487]
[1] 10217.75
>
> ma[486:486]
[1] 9654.484
>
> accuracy(fcastc, ma[487:491], test=1:length(ma[487:491]))
          ME         RMSE          MAE          MPE         MAPE         ACF1    Theil's U
-37.79376842 309.80563169 221.01913411  -0.38996357   2.11724821  -0.05685781   0.95373980
```

Forecasting Combinations

Me

```
> fcastc
[1]  9990.71 10293.34 10554.54 10795.24 11014.49
>
> me[487:487]
[1] 9971.102
>
> me[486:486]
[1] 9479.097
>
> accuracy(fcastc, me[487:491], test=1:length(me[487:491]))
         ME        RMSE         MAE         MPE        MAPE        ACF1   Theil's U
-237.9950413 424.3254152 277.8830721  -2.3331762   2.7169712   0.1419578   1.4003025
```

Final Forecasts

Mi

```
> fcastc
[1]  9663.423  9864.035 10064.648 10265.261 10465.873
```

Mi

```
Point Forecast
  0.001043531
```

Ma

```
> fcastc
[1] 10148.26 10370.51 10575.59 10773.98 10969.77
```

Ma

```
Point Forecast
  0.0006241531
```

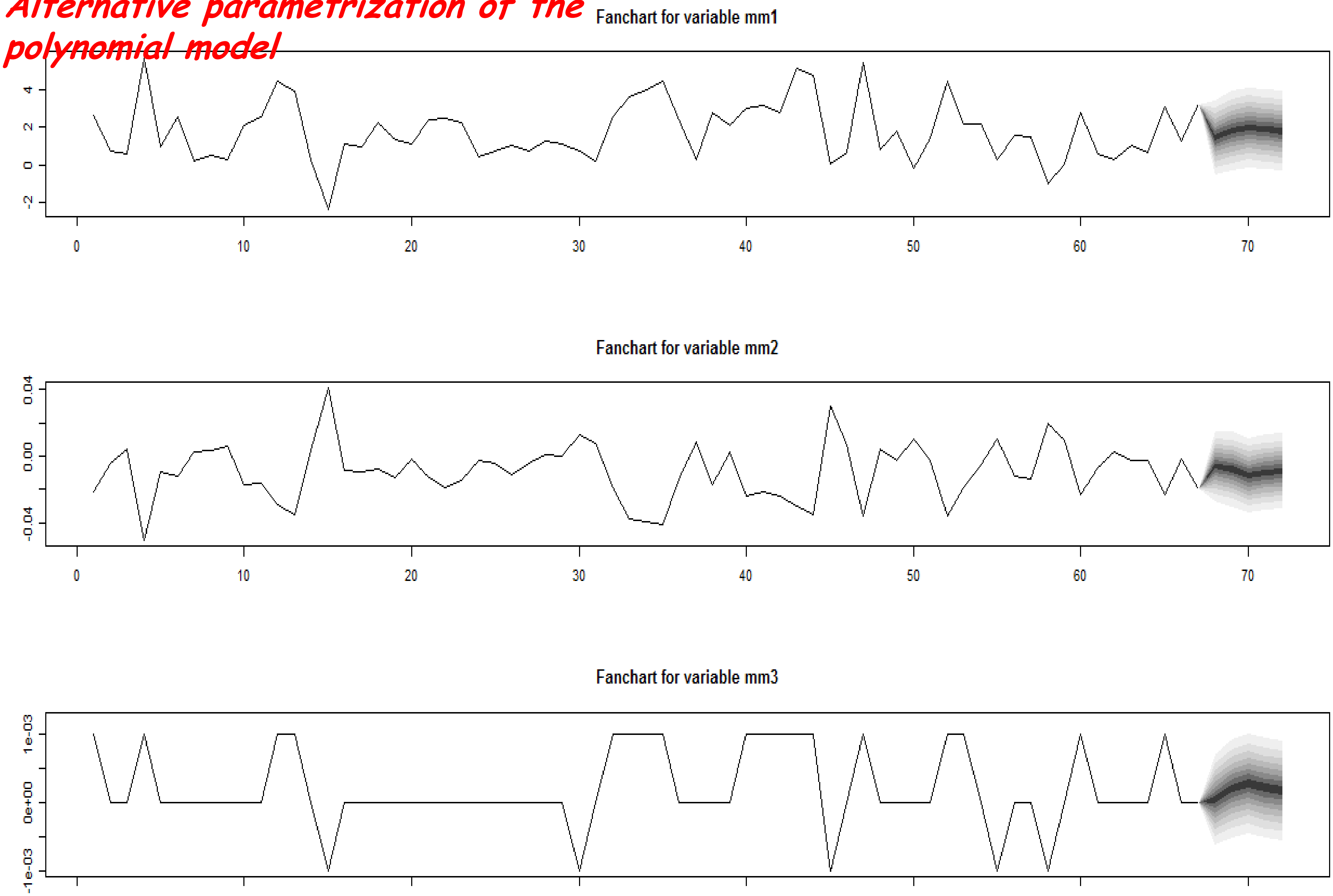
Me

```
> fcastc
[1]  9990.71 10293.34 10554.54 10795.24 11014.49
```

Me

```
Point Forecast
  0.0005747982
```

Alternative parametrization of the polynomial model



Some developments

☞ **Beanplot clustering of different beanplot time series and considering them in a Forecasting Model (see Drago Scepi 2010 presented at Gfkl\Cladag in Karlsruhe).**

☞ **Forecasts Combinations using different forecasting methods**

☞ **Multivariate case: Cointegration (Long run and short run)**

☞ **Beanplot TSFA**

an internal parametrization, where it is crucial to fit adequately (or usefully) the data.

Some References

•Arroyo J. , Gonzales Rivera G., and Matè C. (2009) “Forecasting with Interval and Histogram Data: Some Financial Applications”. Working Paper

•Arroyo J., Matè C. (2009) ” Forecasting Histogram Time Series with K-Nearest Neighbours Methods” International Journal of Forecasting, 25, pp.192-207

•Billard, L., Diday, E. (2006) Symbolic data analysis: conceptual statistics and data mining. Chichester: Wiley & Sons.

•Dacorogna B. et al. (2001) An Introduction of High Frequency Finance. Academic Press.

•Drago C., Scepi G. (2010) “Forecasting by Beanplot Time Series” Electronic Proceedings of Compstat/, Springer Verlag, p.959-967, ISBN 978-3-7908-2603-6

•Drago C., Scepi G. (2010) “Visualizing and exploring high frequency financial data: beanplot time series” accettato su : /New Perspectives in Statistical Modeling and Data Analysis, Springer Series: Studies in Classification, Data Analysis, and Knowledge Organization, Ingrassia, Salvatore; Rocci, Roberto; Vichi, Maurizio (Eds), ISBN: 978-3-642-11362, atteso per novembre 2010

•Engle, R.F, Russel J.R. (2004) “Analysis of High Frequency Financial Data” Working Paper.

•Kampstra, P. (2008) Beanplot: “A Boxplot Alternative for Visual Comparison of Distributions” Journal of Statistical Software Vol. 28, Code Snippet 1, Nov. 2008

•Meijer E., Gilbert P.D. (2005) “Time Series Factor Analysis with an Application to Measuring Money” SOM Research Report, University of Groningen.

•Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. JRSS-B 53, 683-690.

•Yan, B., Zivot G. (2003).Analysis of High-Frequency Financial Data with S-PLUS.Working Paper.