Statistical Significance of Forecasting Methods

An Empirical Evaluation of the Robustness and Interpretability of the MCB, ANOM and Nemenyi Test

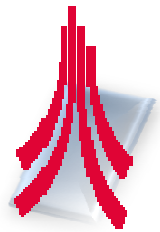
The 32nd Annual International Symposium on Forecasting

Michele Hibon
INSEAD

Sven Crone
Lancaster University Management School

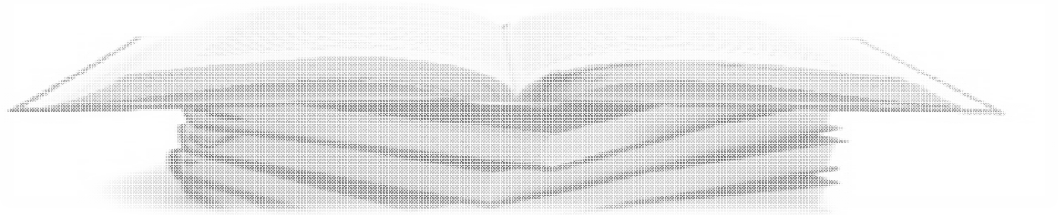
Nikolaos Kourentzes
Lancaster University Management School

INSEAD



LANCASTER UNIVERSITY
www.lancs.ac.uk

Lancaster Centre for Forecasting
www.forecasting-centre.com

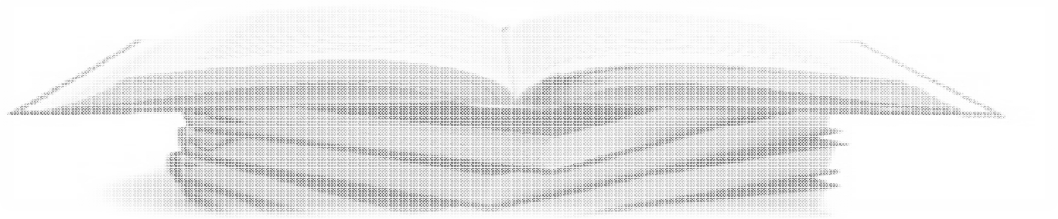


Motivation

Comparison of Models

Principal Question → Which forecasting model is better for X scenario?

- We build/apply several forecasting models → Which one to use?
- Traditionally: Measure forecasting error, rank models & pick first
  - Seen in the M competitions [Makridakis & Hibon, 2000]
  - Model accuracy vs. Complexity
- Statistical tests → Consider uncertainty over measurements
  - Different forecasting errors may not be statistically different!
  - Pair [Diebold & Mariano, 1995, etc] vs. multiple comparison tests [Koning et al, 05, etc]
- What does statistical significance mean? How to make the use of statistical tests straightforward for complex experiments (several models & time series)?

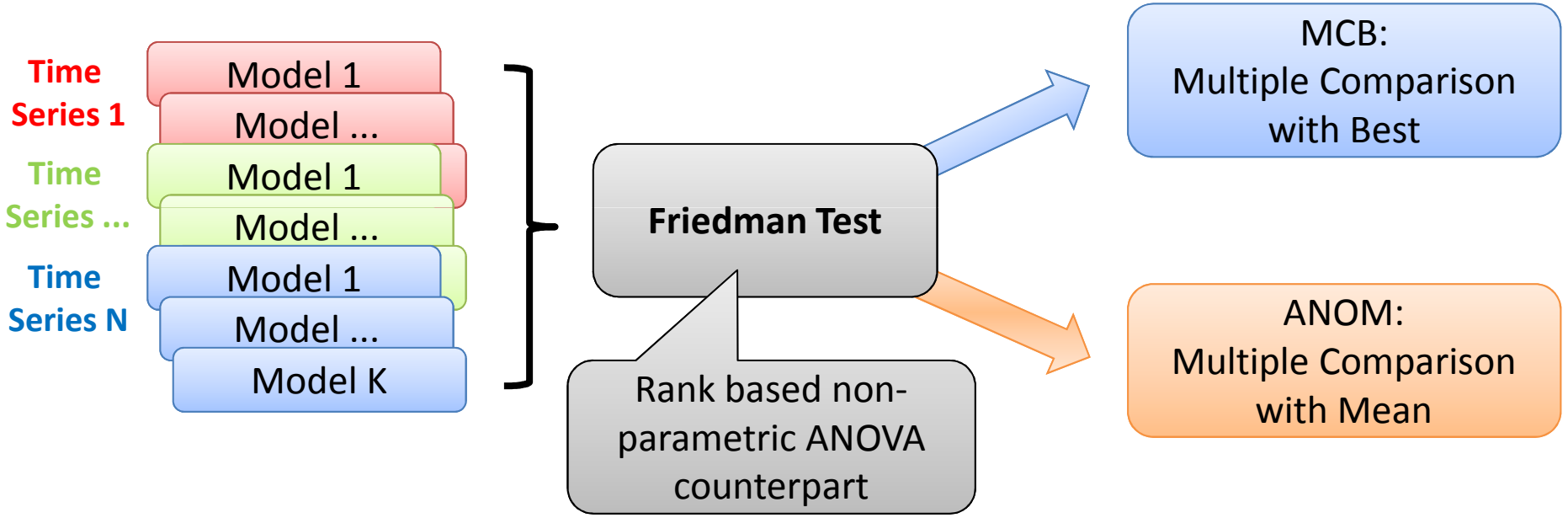


Motivation

Comparison of Models

Multiple Comparison Tests

- Konig et al., 05 presented a series of statistical tests for the M3 competition (based on McDonald & Thomson, 1967, 1972) :



- Demsar, 06 discusses why Friedman in favour of ANOVA and suggests yet another post-hoc test → Nemenyi test [Nemenyi, 63], able to distinguish groups of models

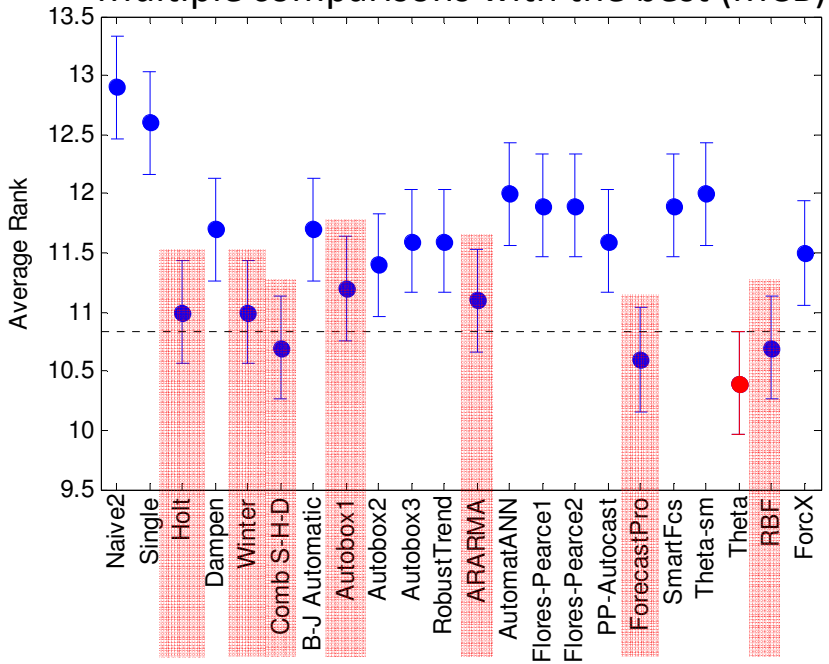
- M3 competition
- 1428 monthly time series
- Forecast horizon: 12

Model	Average Rank
Naive2	12.9
Single	12.6
Holt	11
Dampen	11.7
Winter	11
Comb S-H-D	10.7
B-J automatic	11.7
Autobox1	11.2
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
ARARMA	11.1
AutomatANN	12
Flores-Pearce1	11.9
Flores-Pearce2	11.9
PP-Autocast	11.6
ForecastPro	10.6
SmartFcs	11.9
Theta-sm	12
Theta	10.4
RBF	10.7
ForcX	11.5

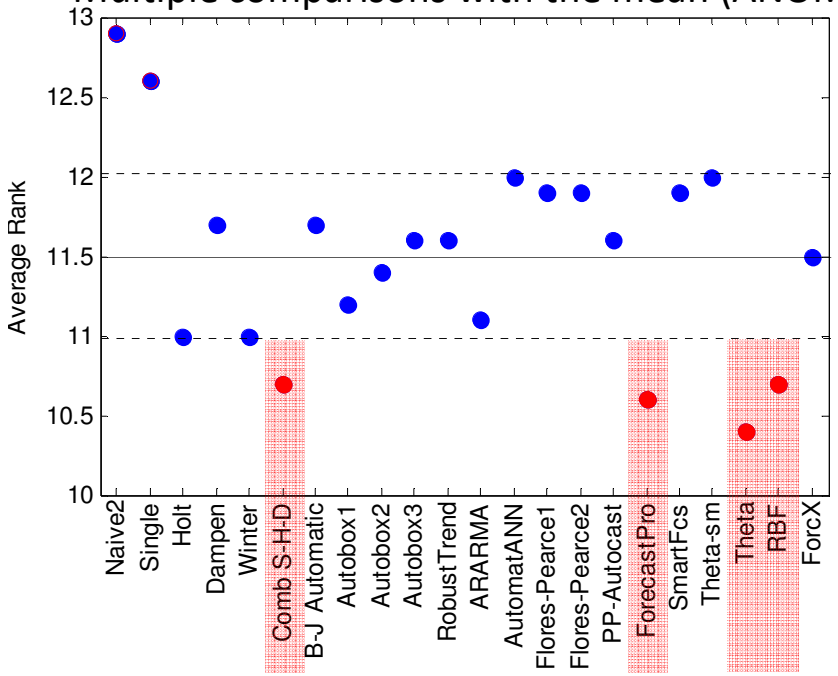
Confidence intervals →
f(# of models, # of time series)

Koning et al, 2005, The M3 competition: Statistical tests of the results, International Journal of Forecasting, 21, 397-409

Multiple comparisons with the best (MCB)

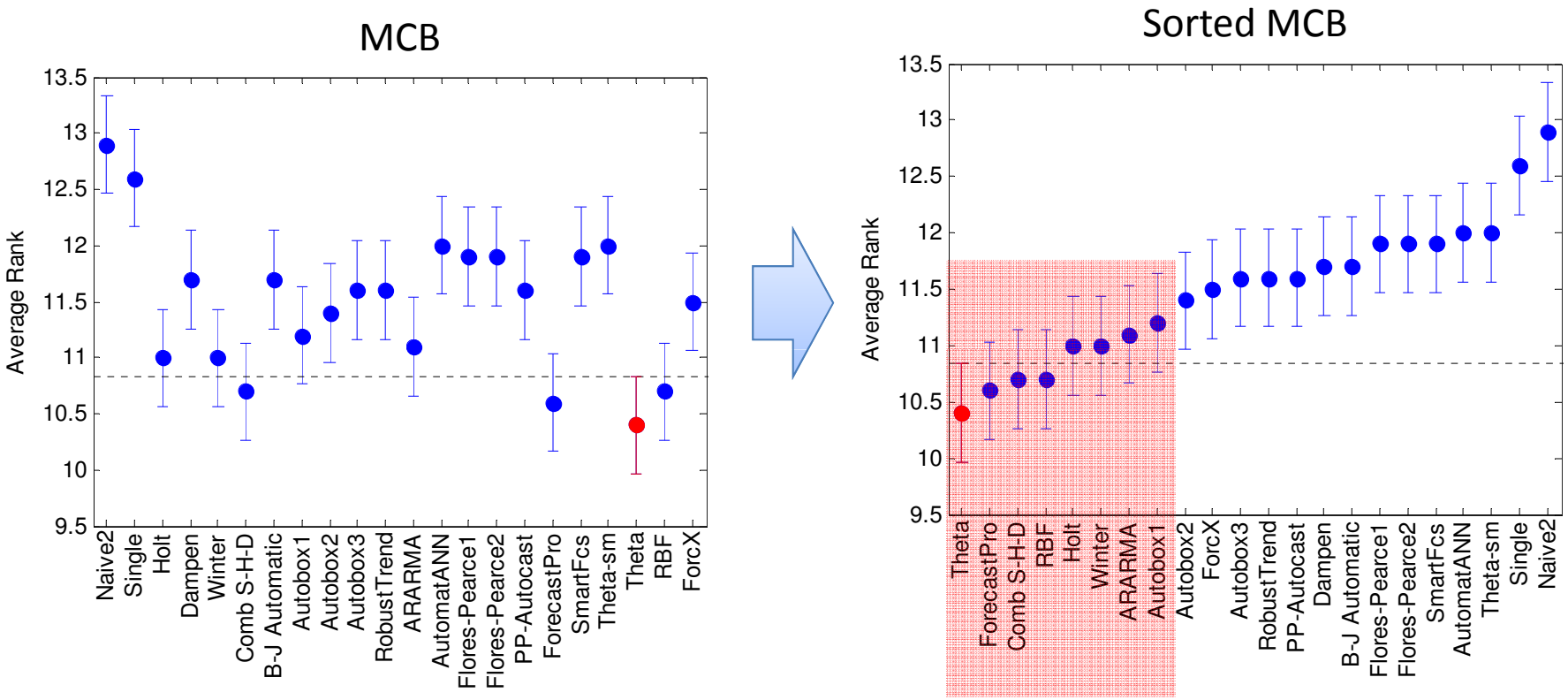


Multiple comparisons with the mean (ANOM)



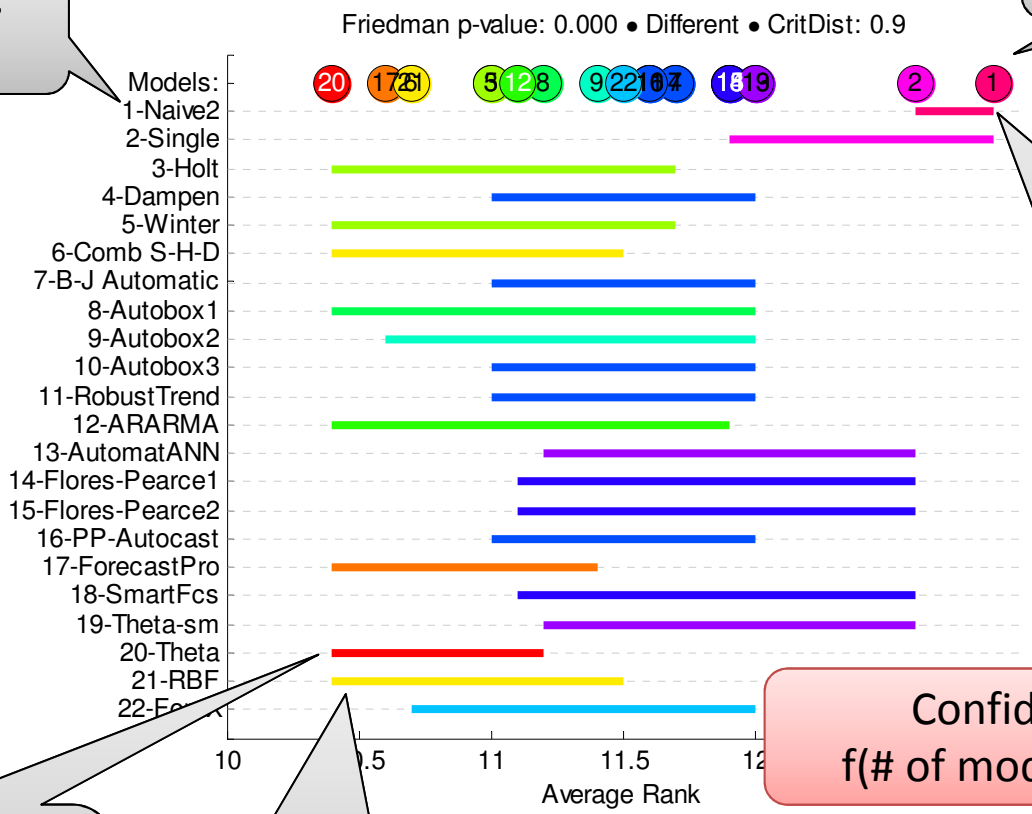
MCB Test

Sorting the models by average rank can increase the readability of MCB



Nemenyi Test

Nemenyi Test → Compares all models and provides groups that do not have significant differences



Model 1: Naive2

Model 1: Naive2

Models 1 and 2 are not significantly different

Joined by line

Confidence intervals → f(# of models, # of time series)

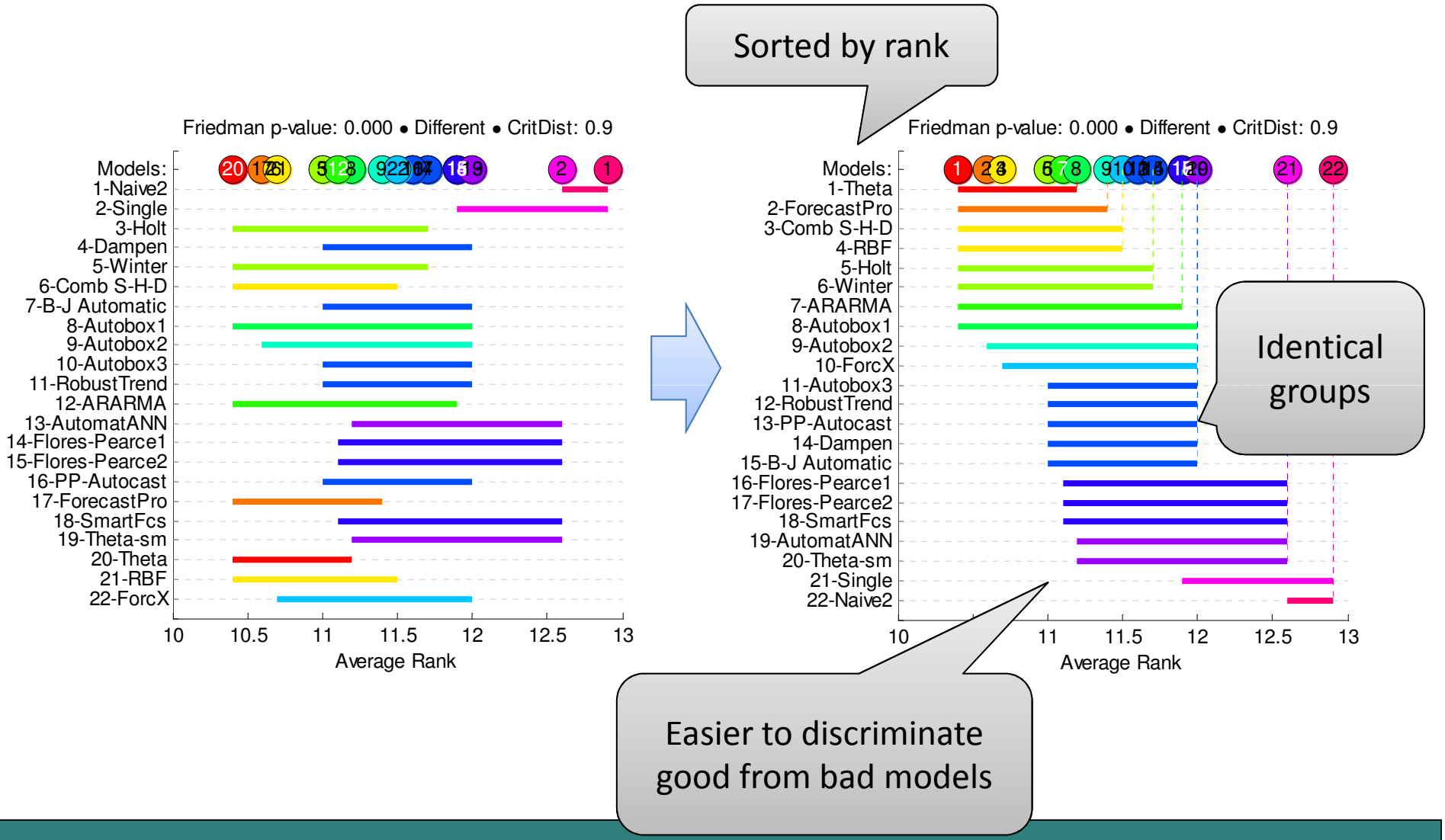
Theta no different than models 17, 6, 21, 3, 5, 12, 8

RBF no different than models 20, 17, 6, 3, 5, 12, 8, 9, 22

Difficult to get clear winner... unless the difference is performance is substantial.
Ranking depends on the model you are measuring from.

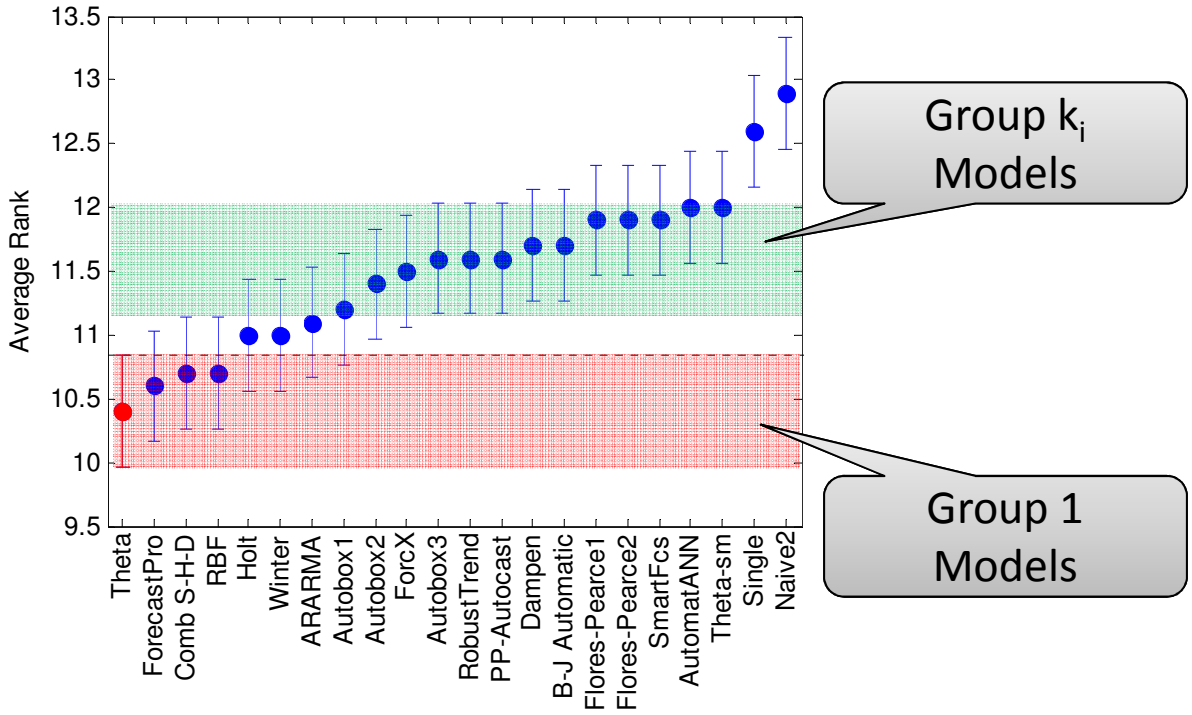
Nemenyi Test

Sorting the Nemenyi test can help in getting a cleaner picture



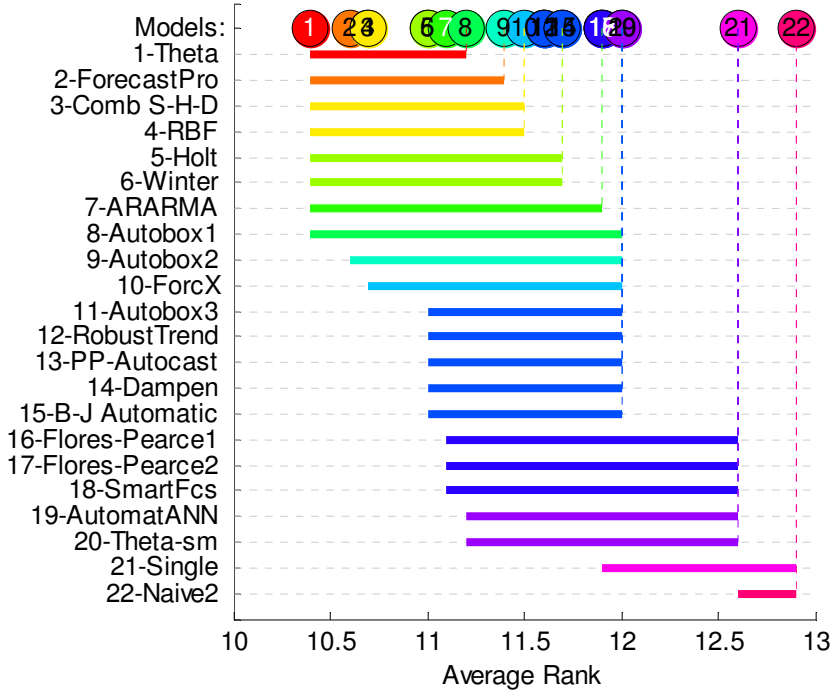
MCB & Nemenyi Test

MCB can become similar to Nemenyi

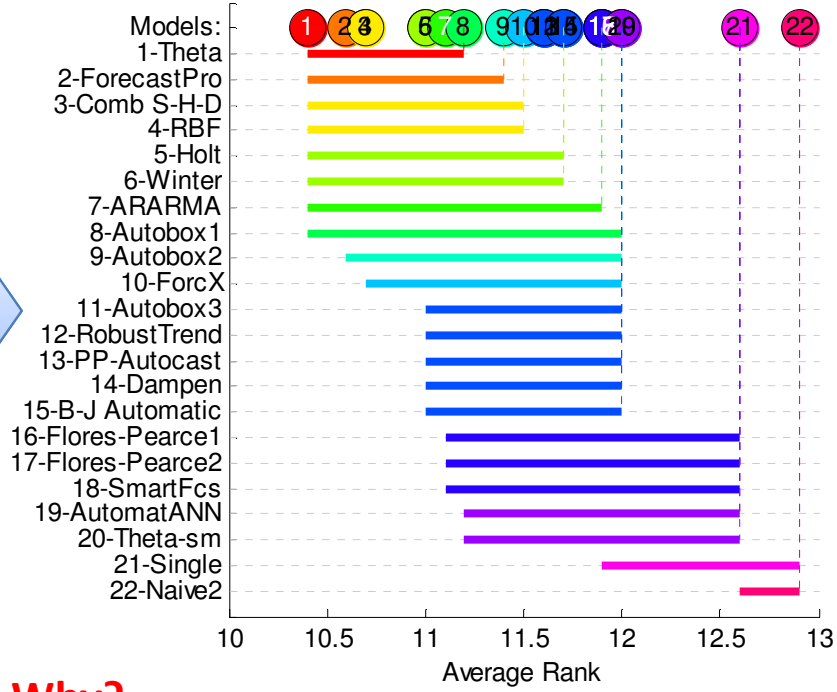


So we can form several groups of models with no significant differences → like Nemenyi tests

MCB



Nemenyi



Identical! Why?

Critical Distance:

$$r_{\alpha,K,N} \approx q_{\alpha,K}\sqrt{\frac{K(K+1)}{12N}}$$

Studentised range for df=∞

For each model: mean rank ± r/2

MCB and Nemenyi is the same test!

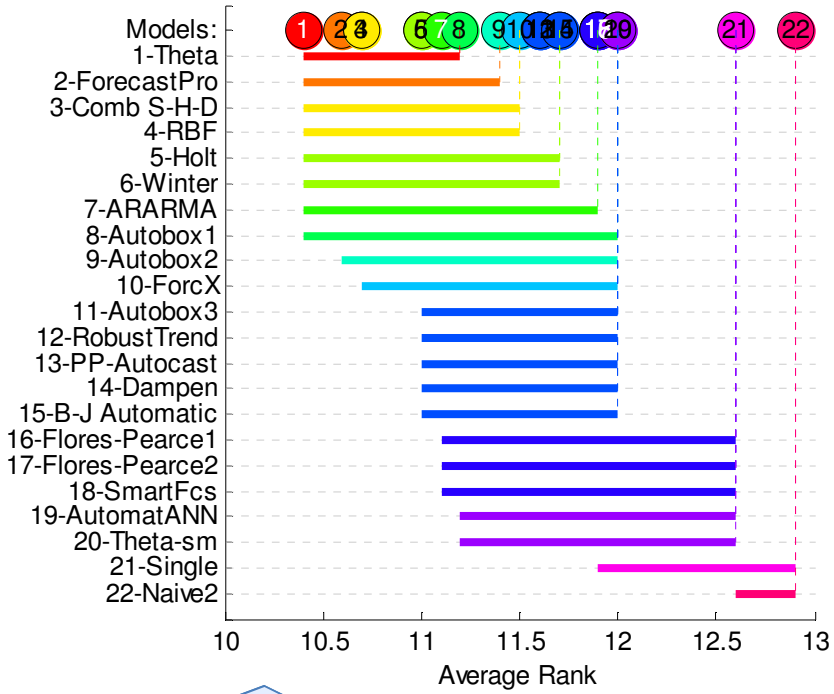
Critical Distance:

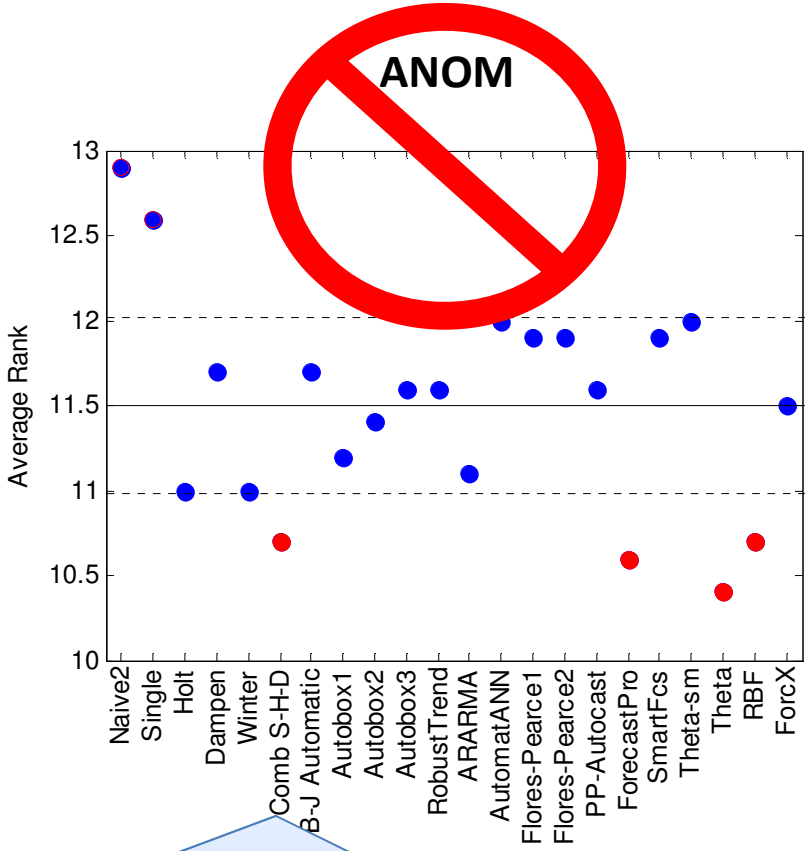
$$r_{\alpha,K,N} \approx \frac{q_{\alpha,K}}{\sqrt{2}}\sqrt{\frac{K(K+1)}{6N}}$$

No differences is models' if mean rank within r

→ Difference between two models: r/2 + r/2 = r

MCB \ Nemenyi

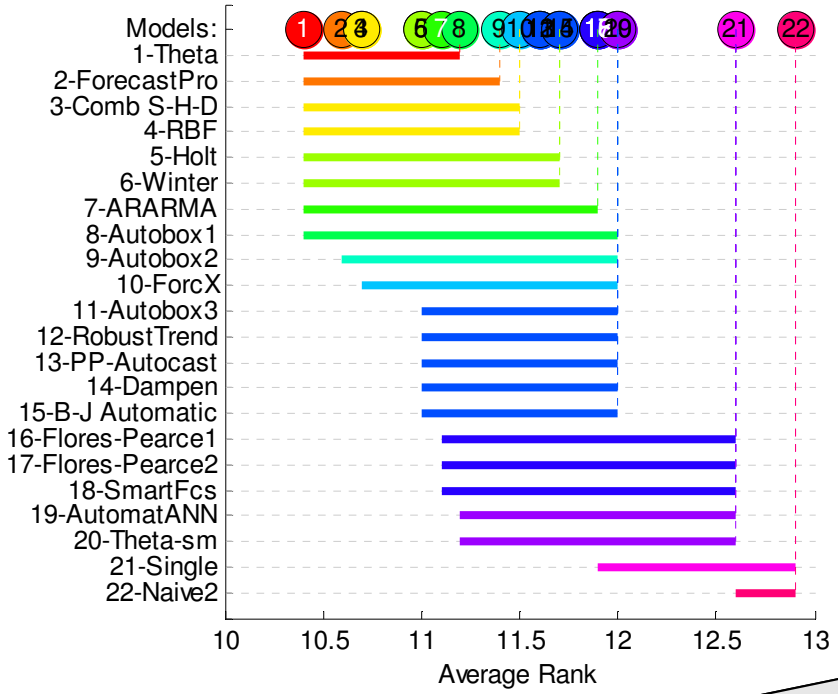


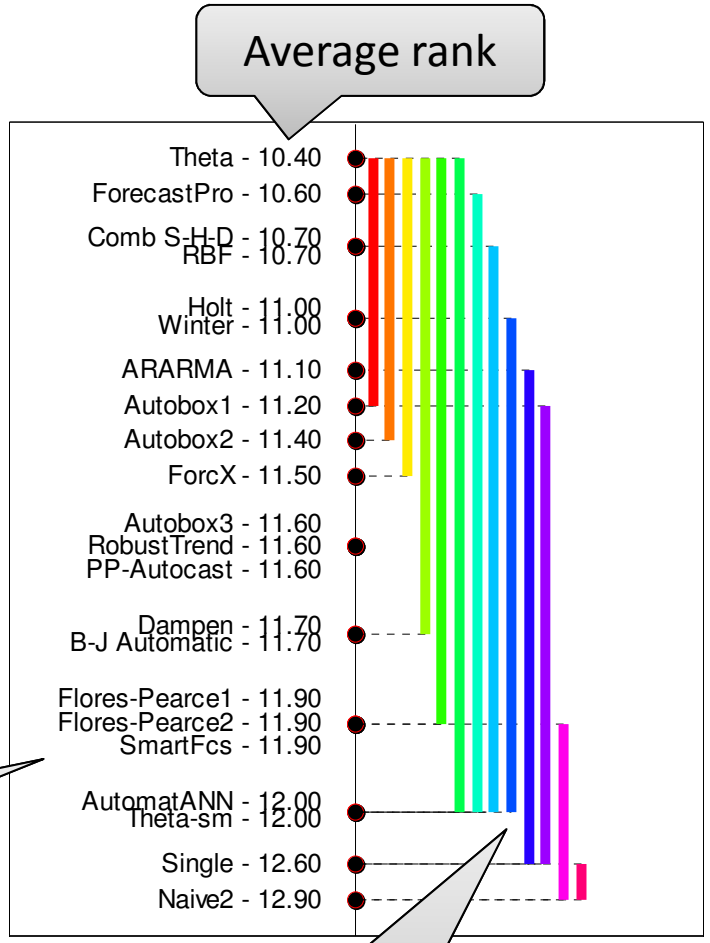


- Much more informative → Model groups
- Simple differences in rank/accuracy not always significant
- No focus on best model only → Simpler & equally accurate model is as good
- M3: Holt/Winter instead of Theta?

- Simple to read... BUT:
- How different are models that deviate significantly from the mean?
- How are the models close to the mean ranked?
- How individual models rank?
- Theta, ForecastPro, Comb S-H-D or RBF?

Yet another visualisation of MCB \ Nemenyi





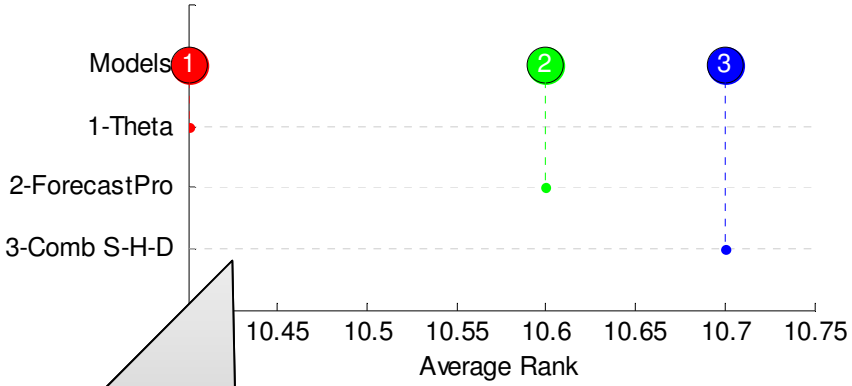
- Allows faster identification of model groups
- Harder to pinpoint how each model ranks (does it matter?)

Same lines combined to one

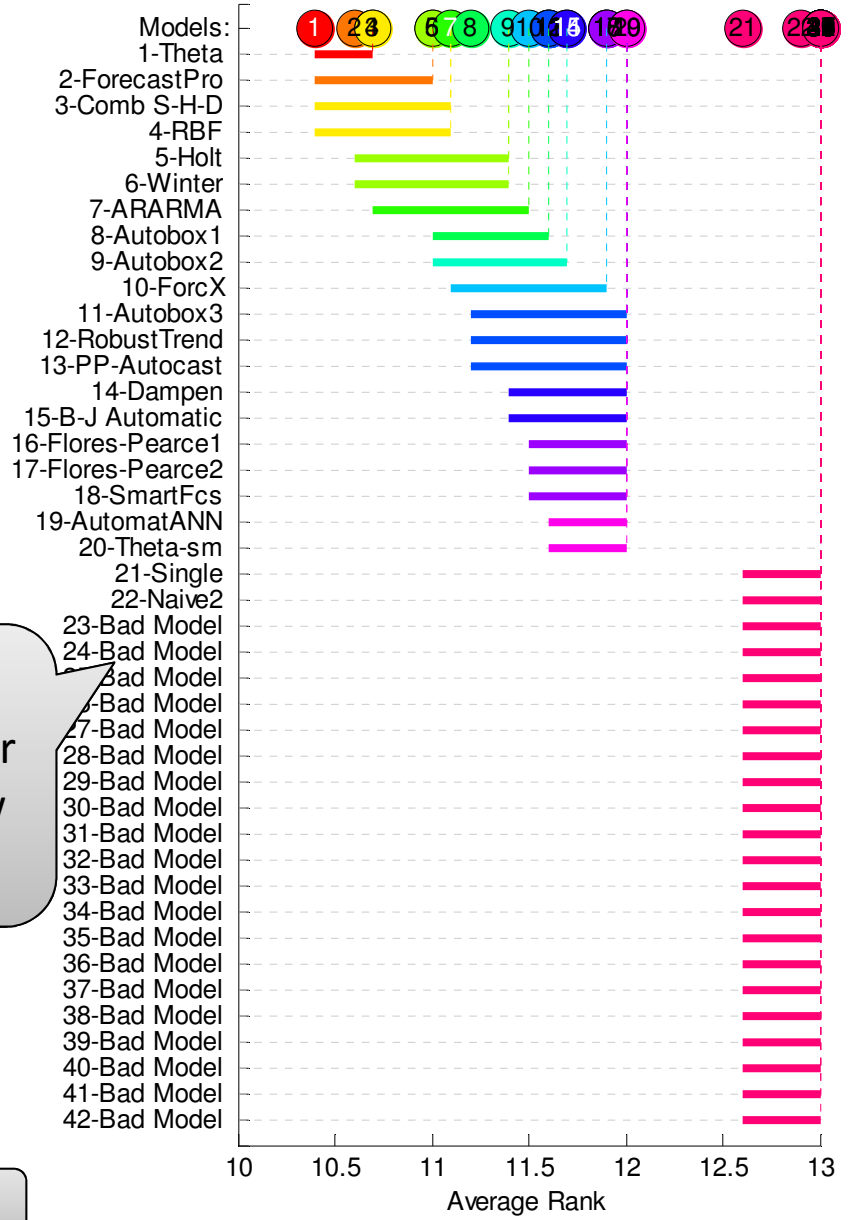
Not always possible to identify single best model or even ranking!

Test Sensitivity

When does MCB/Nemenyi break down?



Pick only 3 top models. Before they were not significantly different, now?



Add a number of “bad models”. Ranks do not change, only number of models. Model groups are now different!

$$r_{\alpha,K,N} \approx q_{\alpha,K}\sqrt{\frac{K(K+1)}{12N}}$$

Number of models

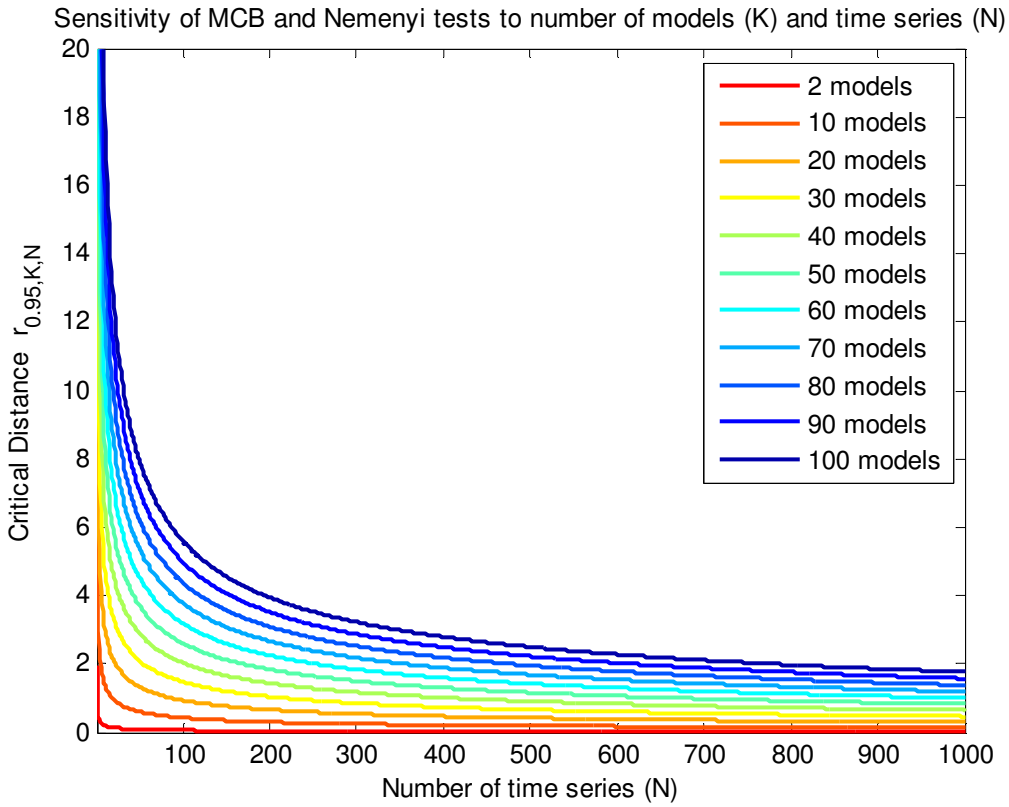
Number of series

Test Sensitivity

When does MCB/Nemenyi break down?

$$r_{\alpha,K,N} \approx q_{\alpha,K}\sqrt{\frac{K(K+1)}{12N}}$$

- Increasing time series → smaller critical distances
- Increasing number of models → larger critical distances
- So, how can you cheat your model to the top?
  → Tweak number of models/time series!

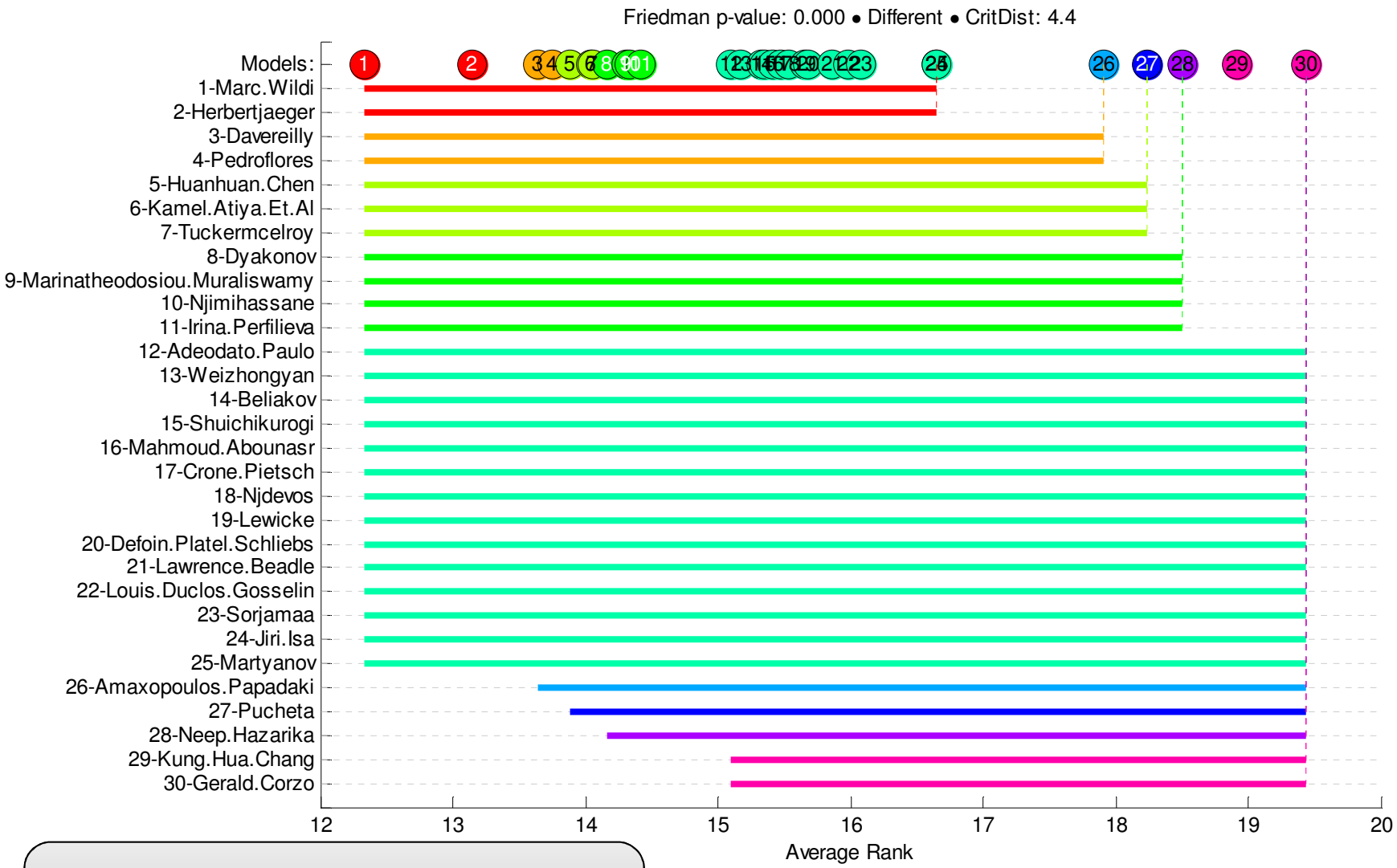


How to design forecasting competitions/experiments?

1. Keep number of models in check! Do not permit multiple similar entries → hides true differences
2. How many time series? Very few will not provide evidence. Too many will make everything significantly different (makes statistical sense, what about practical sense?)
3. M-competitions validity? NN3 competition validity?

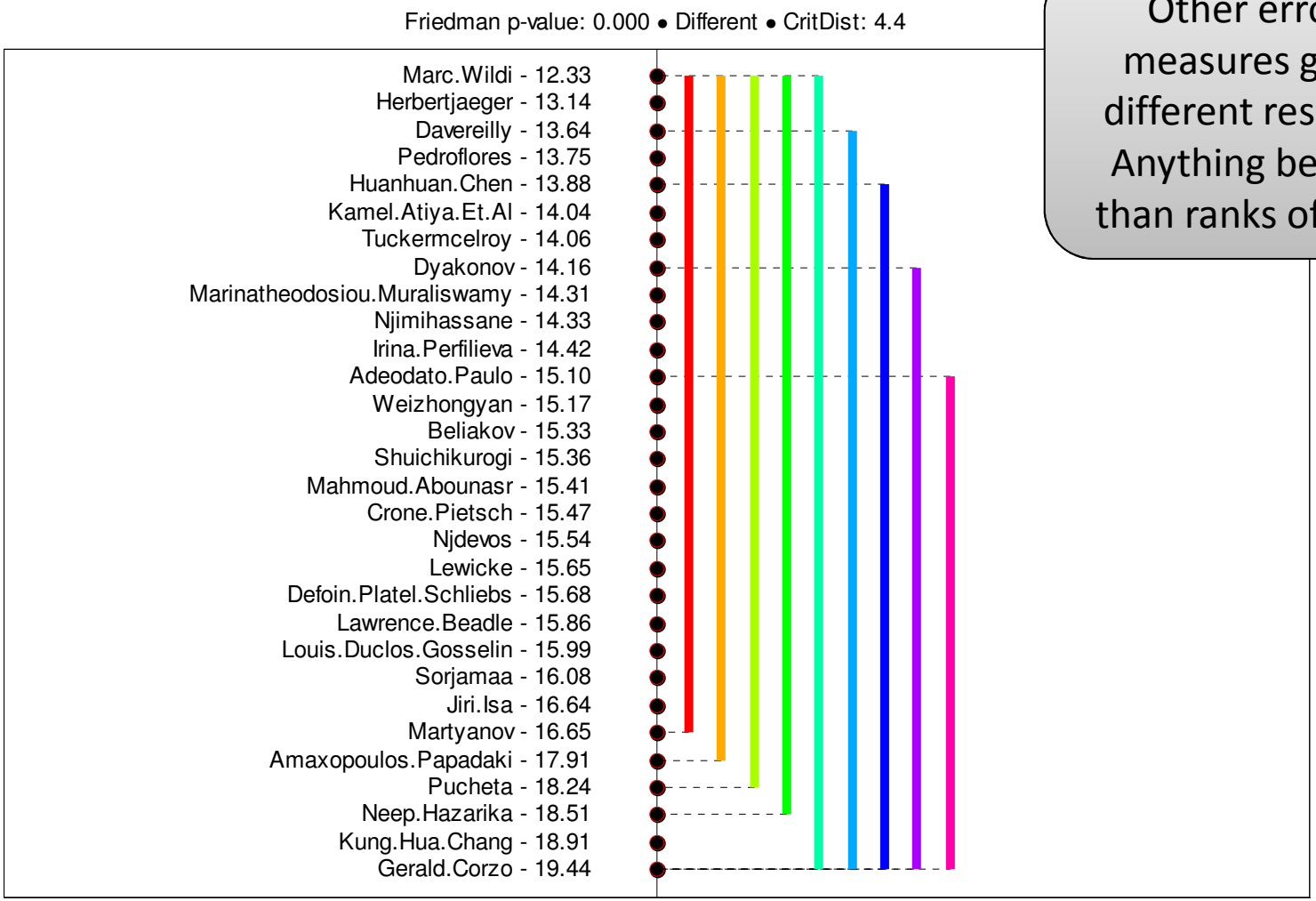
NN3 Complete Dataset (111 time series) Results on Absolute Error

NN3

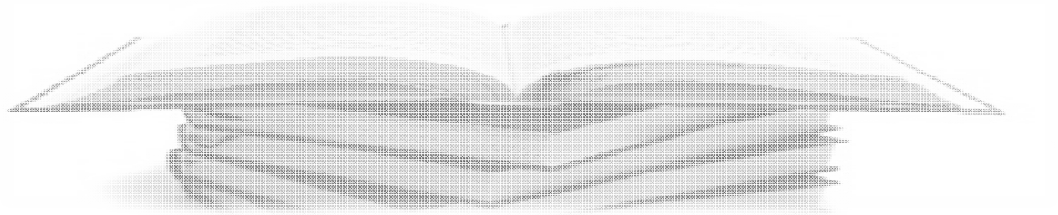


Too many models and too few time series?

NN3 Complete Dataset (111 time series) Results on Absolute Error



Other error measures give different results. Anything better than ranks of AE?



Conclusions

- Model comparison based on statistical tests → Important
- MCB = Nemenyi
- Usefulness of ANOM limited
- Visual tweaks allow easily summarising results across multiple models/time series → replace endless tables...
- Implications for experimental design/forecasting competitions:
  - Tweaking the number of models/time series → Cheating!
  - Choose carefully! But how?

Lancaster University
MANAGEMENT SCHOOL

Thank you for your attention!

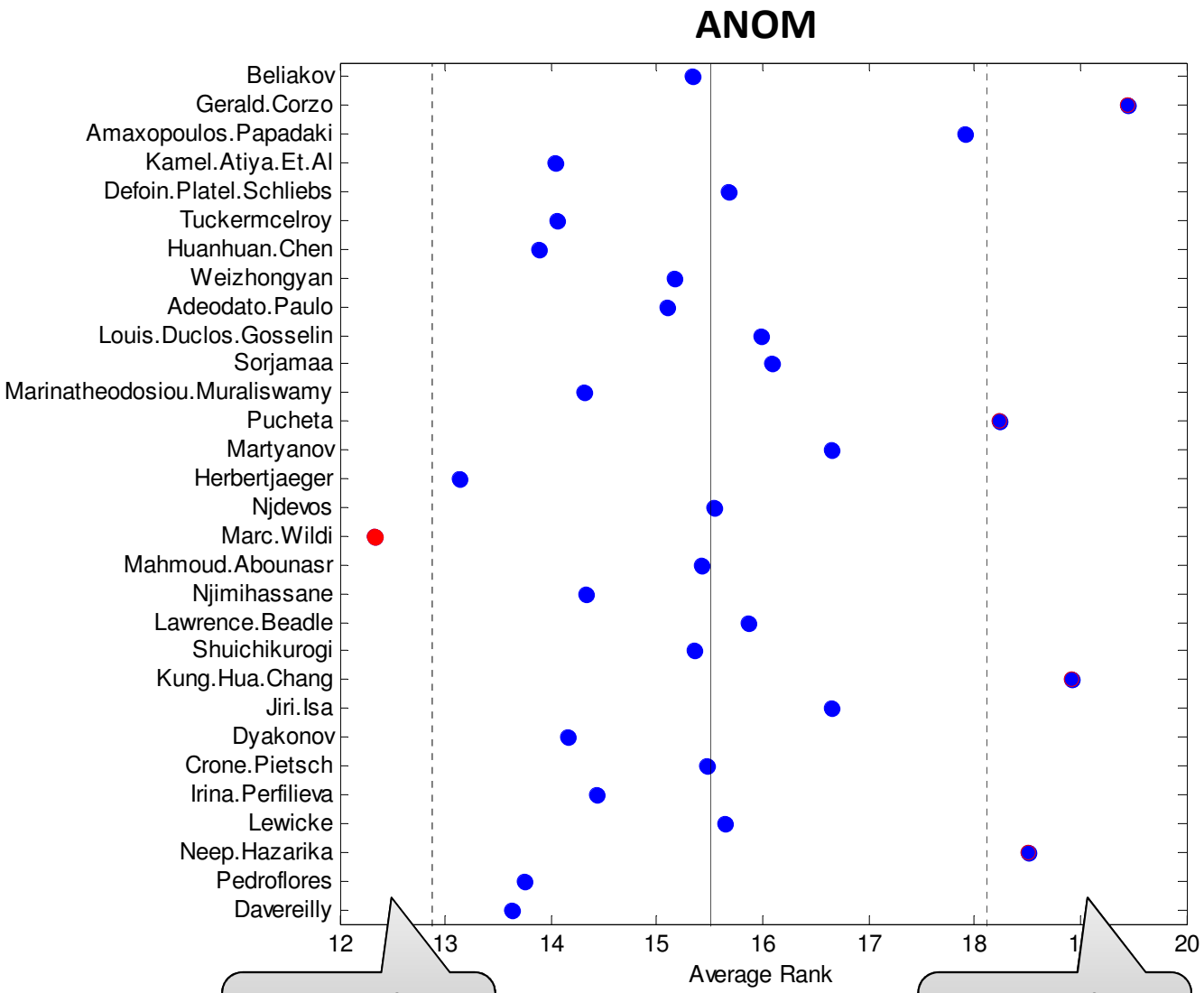
Questions?

Nikolaos Kourentzes

Lancaster University Management School
Centre for Forecasting
Lancaster, LA1 4YX, UK
Tel. +44 (0) 7960271368
email nikolaos@kourentzes.com

NN3 Complete Dataset (111 time series) Results on Absolute Error

NN3



Better than mean rank

Worse than mean rank