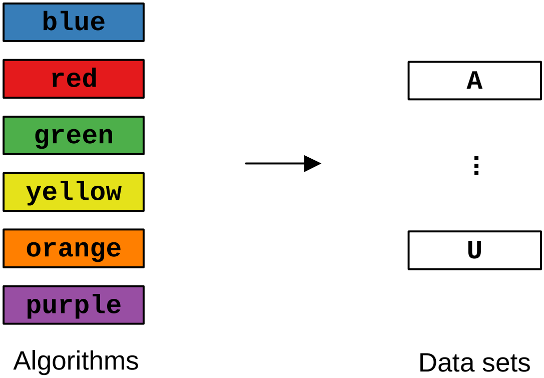
Exploratory and Inferential Analysis of Benchmark Experiments

Manuel J. A. Eugster and Friedrich Leisch

Departement for Statistics
Ludwig-Maximiliams-Universität München

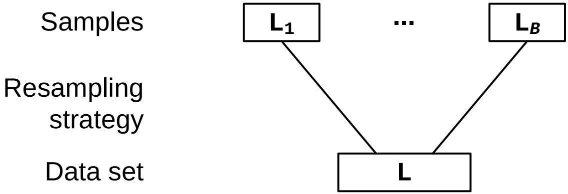
Statistical Computing 2008

“Zehnkampf/Decathlon”



Layers of abstraction

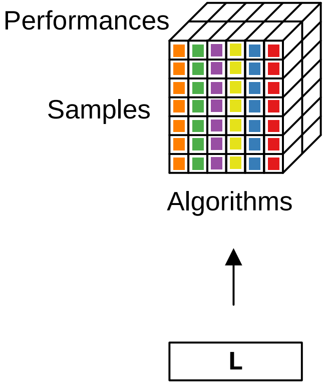
Layer One: Setup



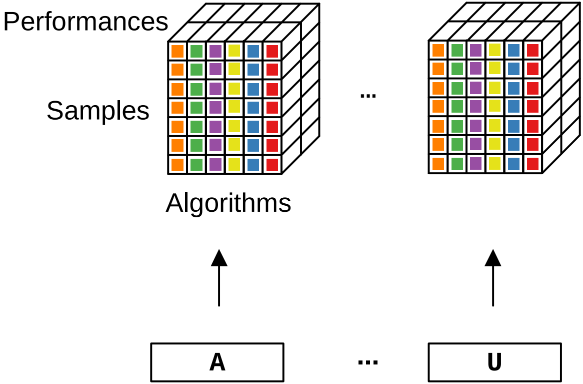
Layer Two: Execution (1)

		L_1	...	L_B
	⋮			
Performances	P_2	0.8		0.7
		8.3		9.1
		2.2		1.9
		12.9		12.3
		0.9		1.1
		1.3		1.3
	P_1	0.020		0.011
		0.219		0.350
		0.372		0.299
		0.014		0.032
		0.386		0.115
		0.299		0.450
Samples		L_1	...	L_B

Layer Two: Execution (2)



Layer Two: Execution (3)



Layer Three: Analysis

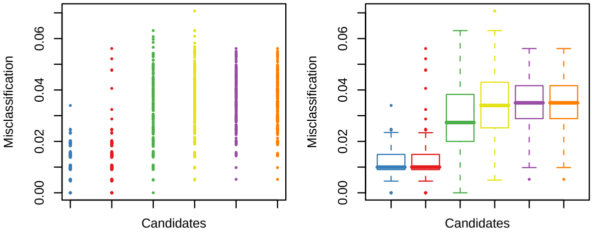
Exploratory: get a better understanding of the benchmark experiment, "dig" for interesting information.

Inferential: test hypotheses of interest, infer a statistically correct order.

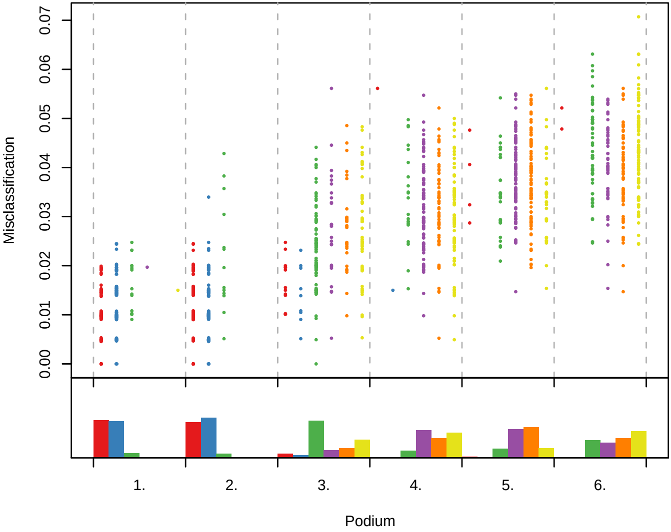
Analyses of benchmark experiments with one data set

Common exploratory tools

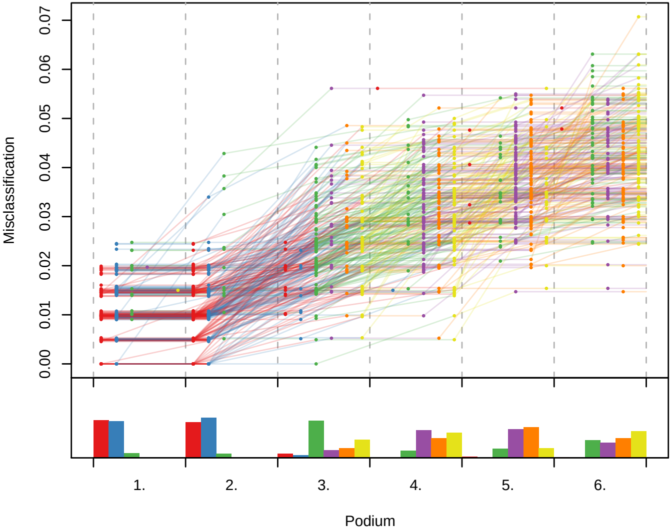
$\phi =$	Mean	SD	Median	Max
`blue`	0.0110	0.0059	0.0100	0.0340
`red`	0.0116	0.0080	0.0100	0.0561
`green`	0.0293	0.0123	0.0273	0.0631
`yellow`	0.0344	0.0118	0.0340	0.0707
`purple`	0.0352	0.0094	0.0350	0.0561
`orange`	0.0353	0.0094	0.0350	0.0561



Benchmark experiment plot



“Full” Benchmark experiment plot



Inferential analysis

Random block design:

$$p_{ij} = \kappa_0 + \kappa_j + b_i + \epsilon_{ij},$$
$$i = 1, \dots, B, j = 1, \dots (K-1),$$

with different assumptions on κ_j, b_i and ϵ_{ij}.

Test problem:

$$H_0 : \ \kappa_1 = \cdots = \kappa_{K-1} = 0,$$
$$H_A : \ \exists j : \ \kappa_j \neq 0,$$

using parametric and non-parametric methods.

Linear mixed effects model

Assumptions:

κ_j fixed effect, b_i random effect,

$$b_i \sim N(0, \sigma_b^2), \epsilon_{ij} \sim N(0, \sigma^2).$$

Test problem:

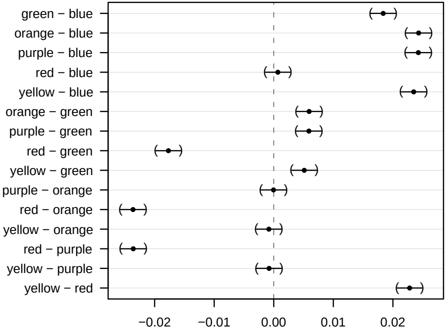
Pairwise comparisons with Tukey contrasts.

Pairwise comparisons based on LME

```
 General Linear Hypotheses

Multiple Comparisons of Means:
Tukey Contrasts

Linear Hypotheses:
                        Estimate
green - blue == 0      1.837e-02
orange - blue == 0     2.431e-02
purple - blue == 0     2.427e-02
red - blue == 0        6.863e-04
yellow - blue == 0     2.349e-02
orange - green == 0    5.941e-03
purple - green == 0    5.899e-03
red - green == 0      -1.769e-02
yellow - green == 0    5.121e-03
purple - orange == 0 -4.188e-05
red - orange == 0     -2.363e-02
yellow - orange == 0 -8.202e-04
red - purple == 0     -2.359e-02
yellow - purple == 0 -7.783e-04
yellow - red == 0      2.281e-02
```



Order relation and toplogical sort

In case of a significant difference between two algorithms we define a strict total order $<$, otherwise the algorithms are $\approx$-related.

Pairwise orders:

$\texttt{red} \approx \texttt{blue}$, $\texttt{purple} \approx \texttt{orange}$, $\texttt{blue} < \texttt{green}$, ...

Topological sort:

$$\texttt{blue} \approx \texttt{red} < \texttt{green} < \texttt{orange} \approx \texttt{purple} \approx \texttt{yellow}$$

Overall order

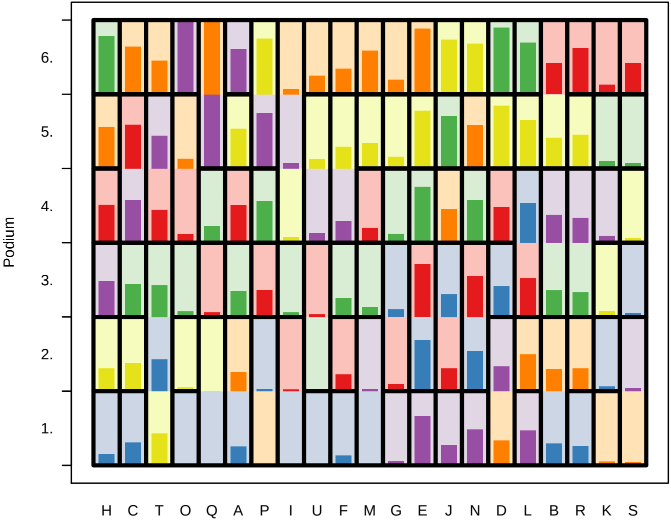
Performance measures P_i:

$$\begin{aligned} \text{Mcl}: &\ \texttt{blue} \approx \texttt{red} < \texttt{green} < \texttt{orange} \approx \texttt{purple} \approx \texttt{yellow} \\ \text{Time}: &\ \texttt{red} < \texttt{purple} < \texttt{orange} < \texttt{yellow} < \texttt{green} < \texttt{blue} \end{aligned}$$

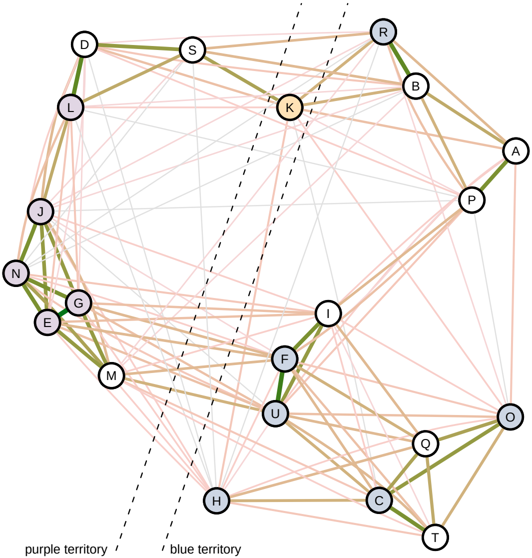
Overall order: Hierarchical order*, Consensus ranking*

Analyses of benchmark experiments with more than one data set

Benchmark survey plot



Benchmark survey graph



Further formal analyses

Consensus: overall order based on the set of order relations.*

Inference: model the design with two experimental factors, their interactions and blocking factors at two levels.*

Overall: sum up order relations based on different data sets and different performance measures.*

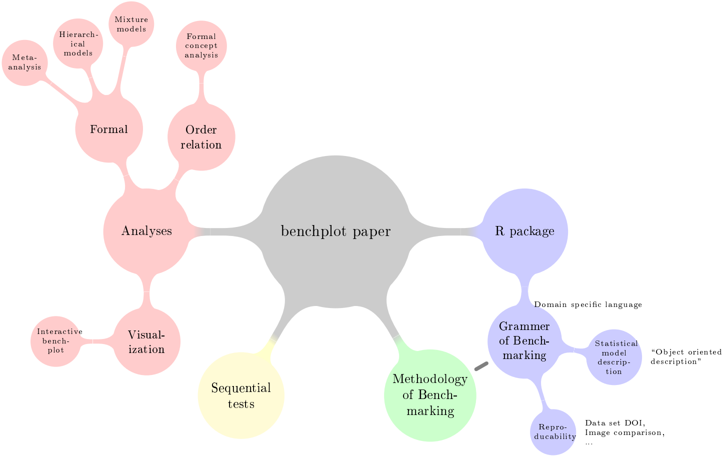
Statistically correct order

Algorithms $\{\texttt{blue}, \texttt{green}, \texttt{orange}, \texttt{red}, \texttt{purple}, \texttt{yellow}\}$, data sets $\{\texttt{A}, \ldots, \texttt{U}\}$, performance measures $P_1 =$ *misclassification*, $P_2 =$ *computation time*:

$$\texttt{blue} < \texttt{red} \approx \texttt{orange} \approx \texttt{green} < \texttt{yellow} < \texttt{purple}$$

Perspective

Goals and future work



References

Bench Plot and Mixed Effects Models: First steps toward a comprehensiv benchmark analysis toolbox.
Manuel J. A. Eugster and Friedrich Leisch. Technical Report 26, LMU Munich. Accepted for the Compstat 2008-Proceedings in Computational Statistics.

(*) *Exploratory and Inferential Analysis of Benchmark Experiments*.
Manuel J. A. Eugster, Torsten Hothorn and Friedrich Leisch. Technical Report 30, LMU Munich.

`http://www.statistik.lmu.de/~eugster`