Validációs paraméterek mintaérettől való függésének vizsgálata

Kovács Dániel, Király Péter, Tóth Gergely

ELTE, Kémiai Intézet

2019. június 6.

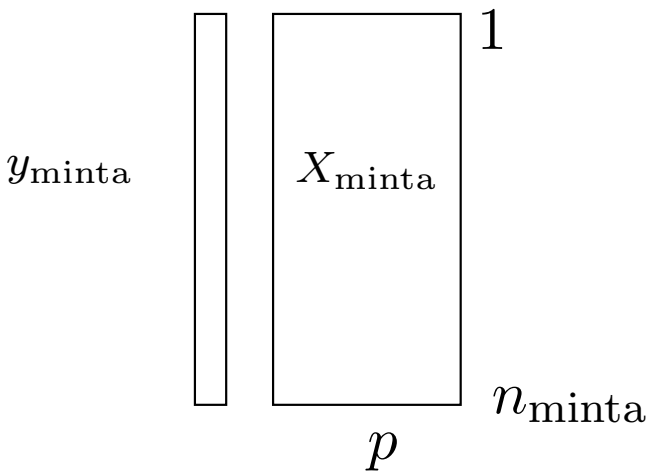
Kivonat

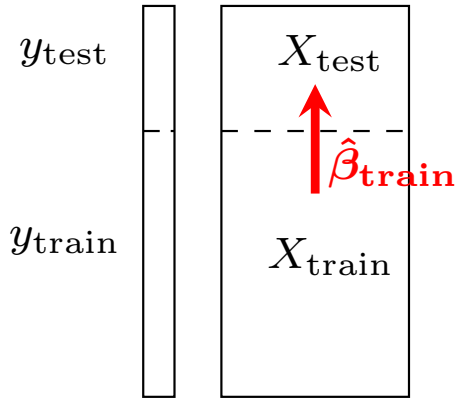
- Validációs paraméterek (VPk) mintamérettől való függése
  - összevetése gyenge és jó illeszkedésű lineáris modelleken
  - R^2 scrambling, randomizációs változatok
  - Q^2 leave one out (LOO) - leave many out (LMO)
  - modellek Roy-Ojha diagramon (Q^2_{F2})
  - VPk rangkorrelációja
- Összegzés

Belső validáció

Külső validáció

$$y = X\beta + e_i$$





illeszkedés jósága, robusztusság

$$r_i = (y_i - \hat{y}_i)$$

prediktivitás

$$r_i = \left(y_{i,\text{test}} - \hat{y}_{i,\text{test}}\left(\hat{\beta}_{\text{train}}\right)\right)$$

belső VPk

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$R^2_{\mathrm{adj}} = \left(1 + \frac{p-1}{n-p}\right) R^2 - \left(\frac{p-1}{n-p}\right)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\mathrm{PRESS}}{\mathrm{TSS}}$$

$$\mathrm{CCC} = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})\left(\hat{y}_i - \bar{\hat{y}}\right)}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2 + n\left(\bar{y} - \bar{\hat{y}}\right)^2} \quad \cdots$$

külső VPk

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n,\mathrm{test}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n,\mathrm{test}}(y_i - \bar{y}_{\mathrm{train}})^2}$$

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n,\mathrm{test}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n,\mathrm{test}}(y_i - \bar{y}_{\mathrm{test}})^2}$$

$$Q^2_{F3} = 1 - \frac{\sum_{i=1}^{n,\mathrm{test}}(y_i - \hat{y}_i)^2/n_{\mathrm{test}}}{\sum_{i=1}^{n,\mathrm{train}}(y_i - \bar{y}_{\mathrm{train}})^2/n_{\mathrm{train}}}$$

$\vdots$

VPk mintaméret függése - Valinear(R)

Adatsorok

Beton nyomószilárdság sor

Kombinált ciklusú erőmű sor

- modellezés *többváltozós lineáris regresszióval* (TLR)

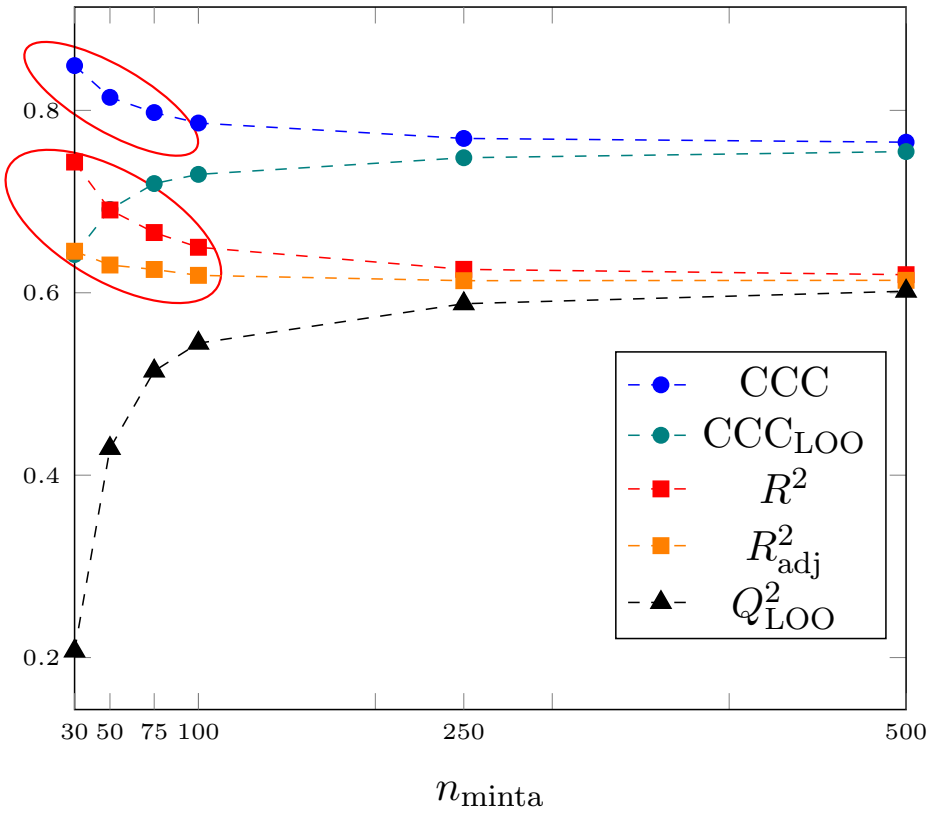
adatsor	Beton	Erőmű
# függő változó	1	1
# független változó	8	4
# megfigyelés	1030	9658
TLR alkalmasság	**gyenge**	**jó**

Mintaméret függés - feladatok

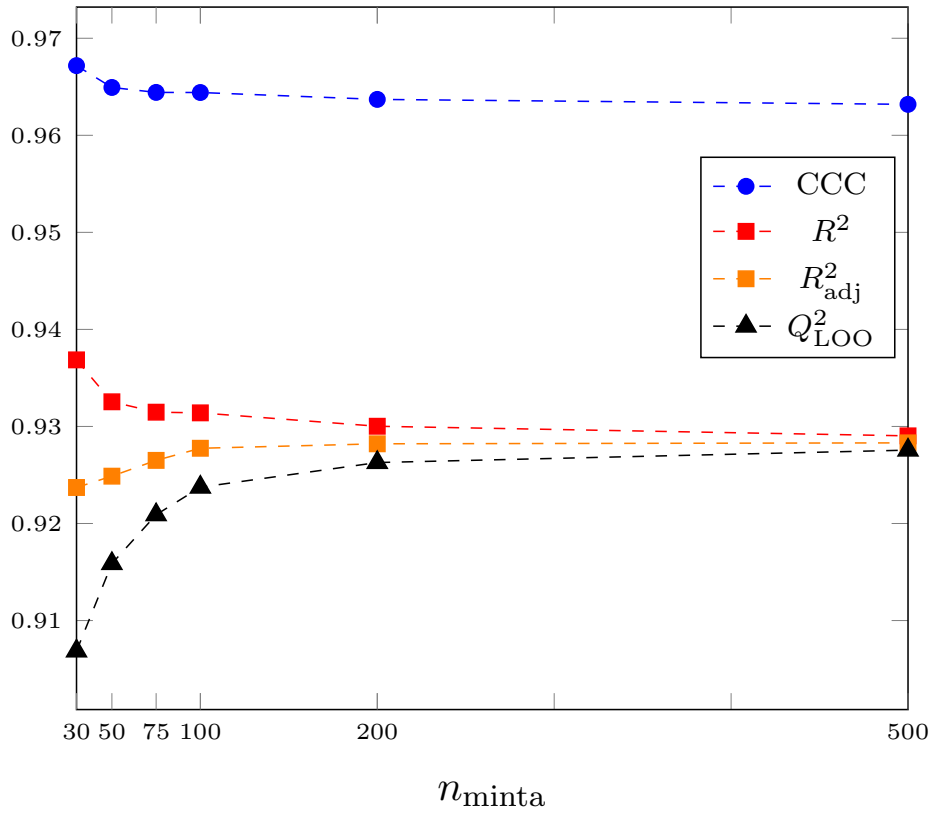
- random mintavételezés a populációból
- minta *training/test* felbontása külső validációhoz: 80/20
- lineáris modellezés, *validációs paraméterek* számolása

↻ fentiek ismétlése $1000\times$, $n_{\text{minta}} = (30, 50, 75, 100, 250, 500)$, VPk átlagolása

Beton

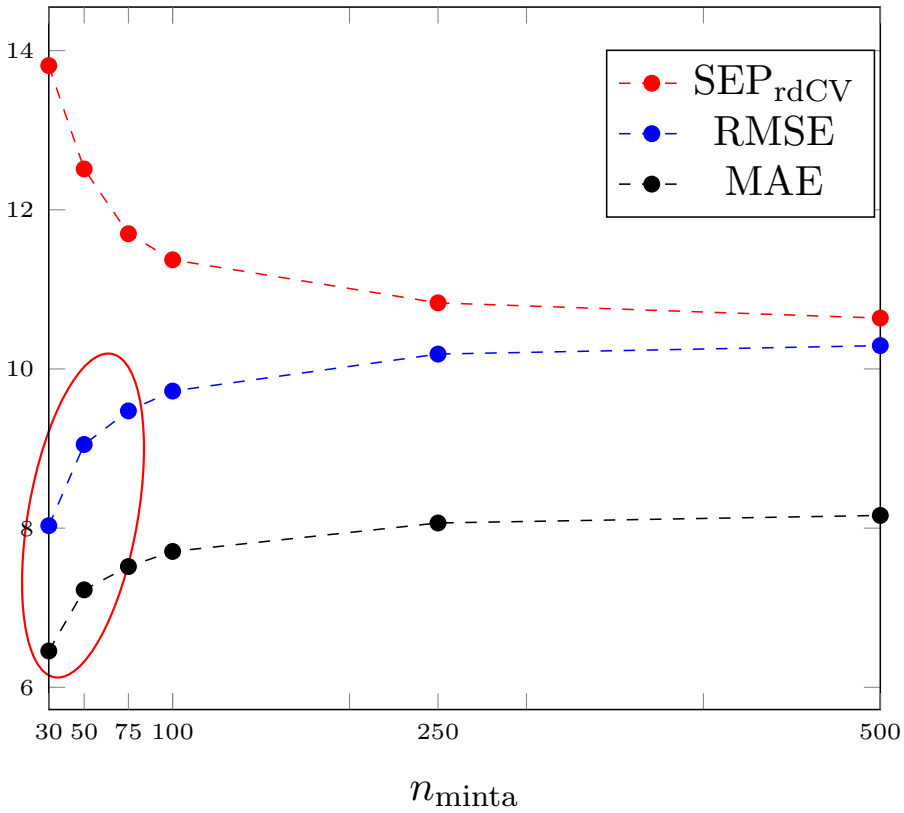


Erőmű

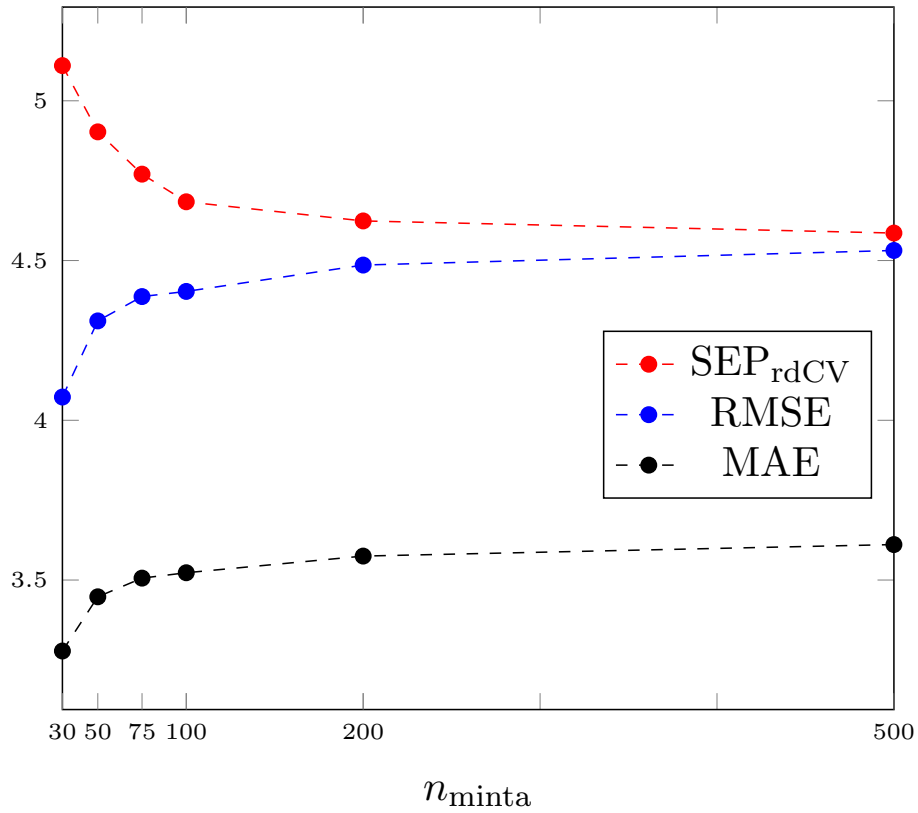


Félrevezető: Kisebb modell a jobb?

Beton

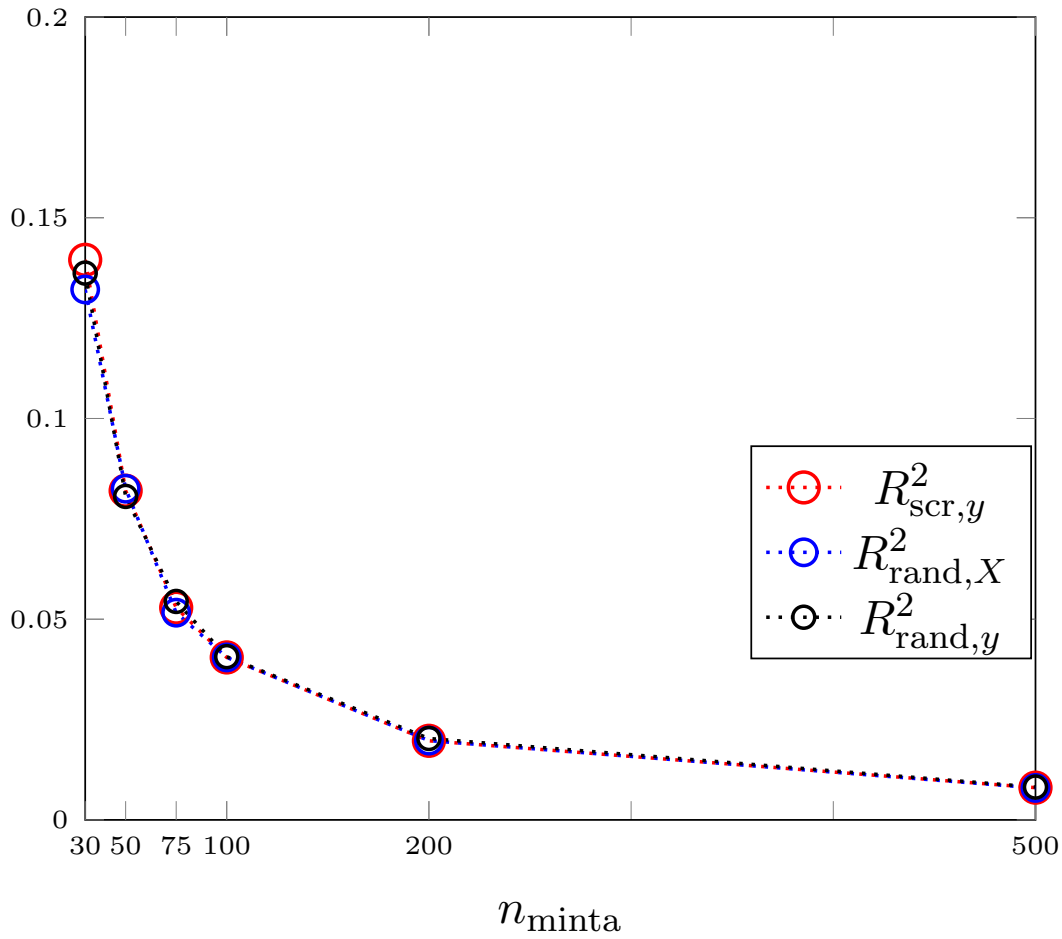


Erőmű



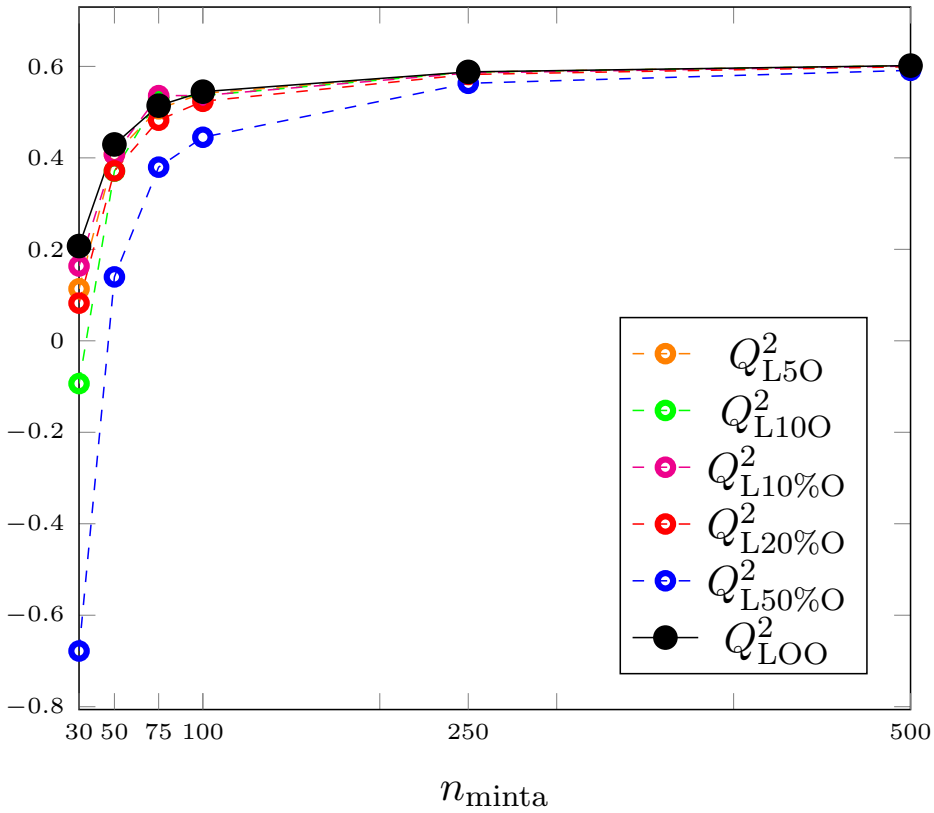
Félrevezető: Kisebb modell a jobb?

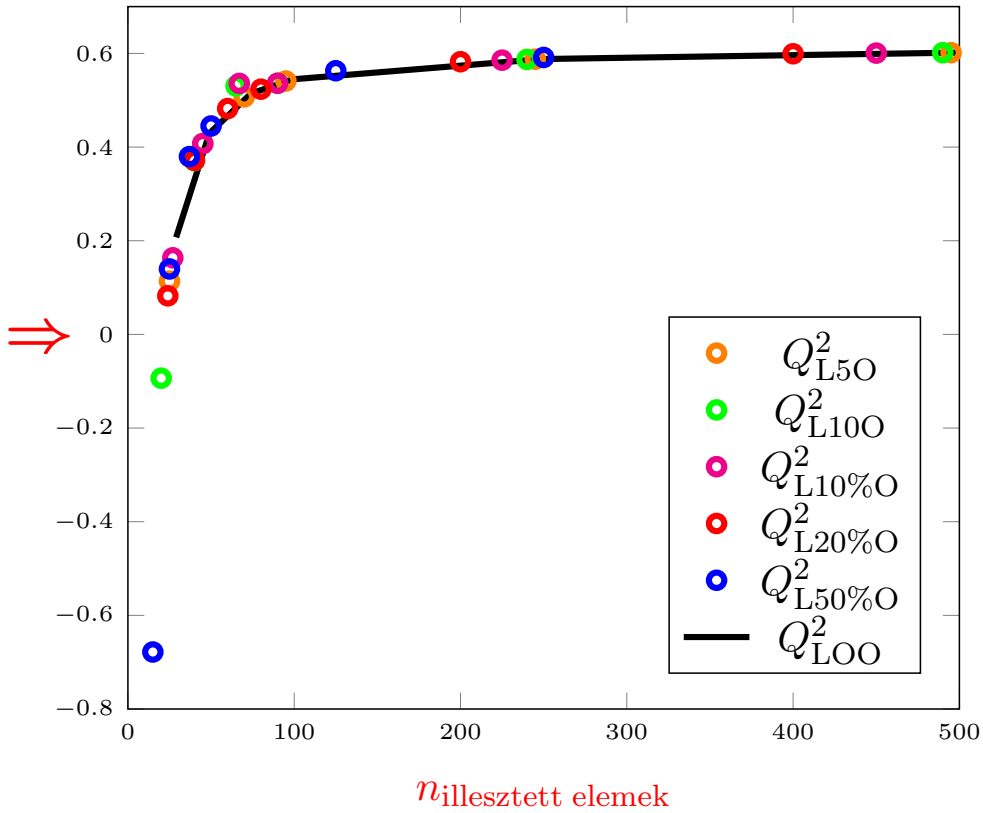
Belső VPk - scrambling, randomizáció



R^2 scrambling számolható a legkönnyebben

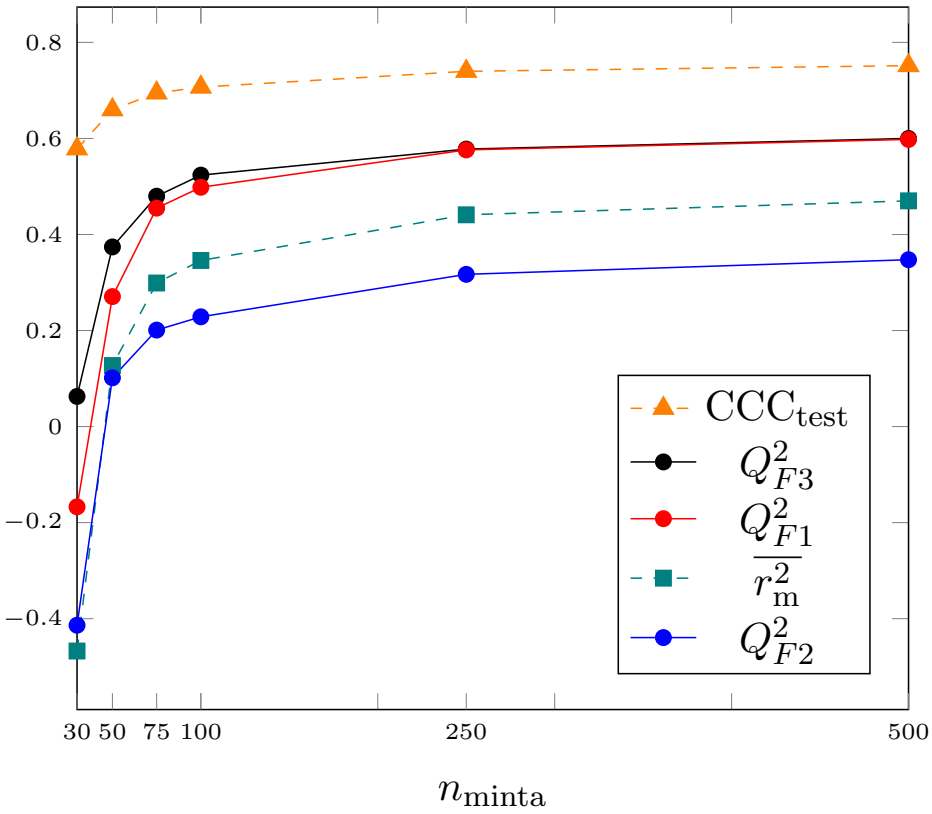
Belső VPk - Leave many out



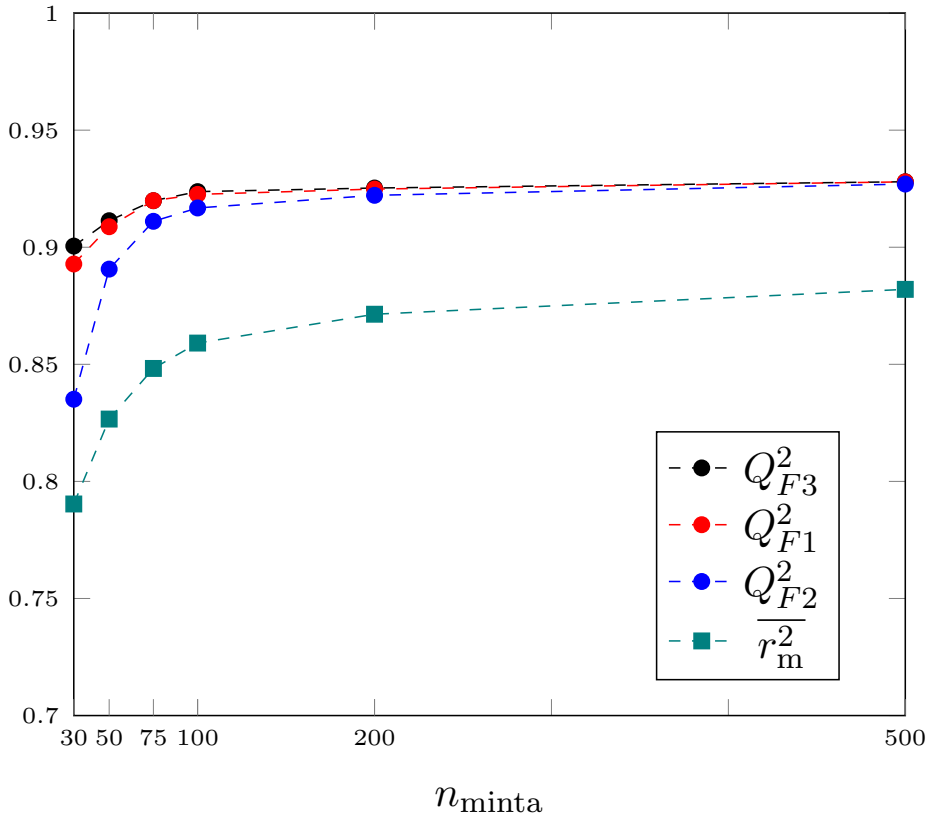


Q^2_{LOO} számolása a legegyszerűbb

Beton



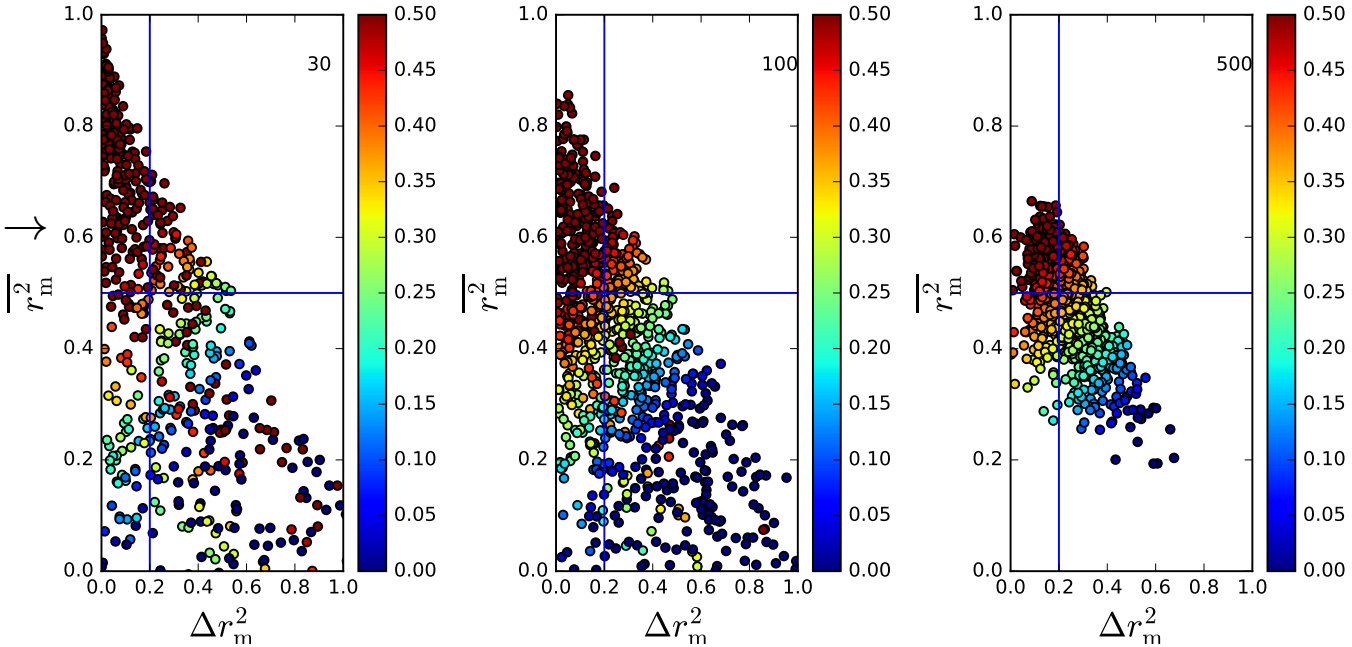
Erőmű



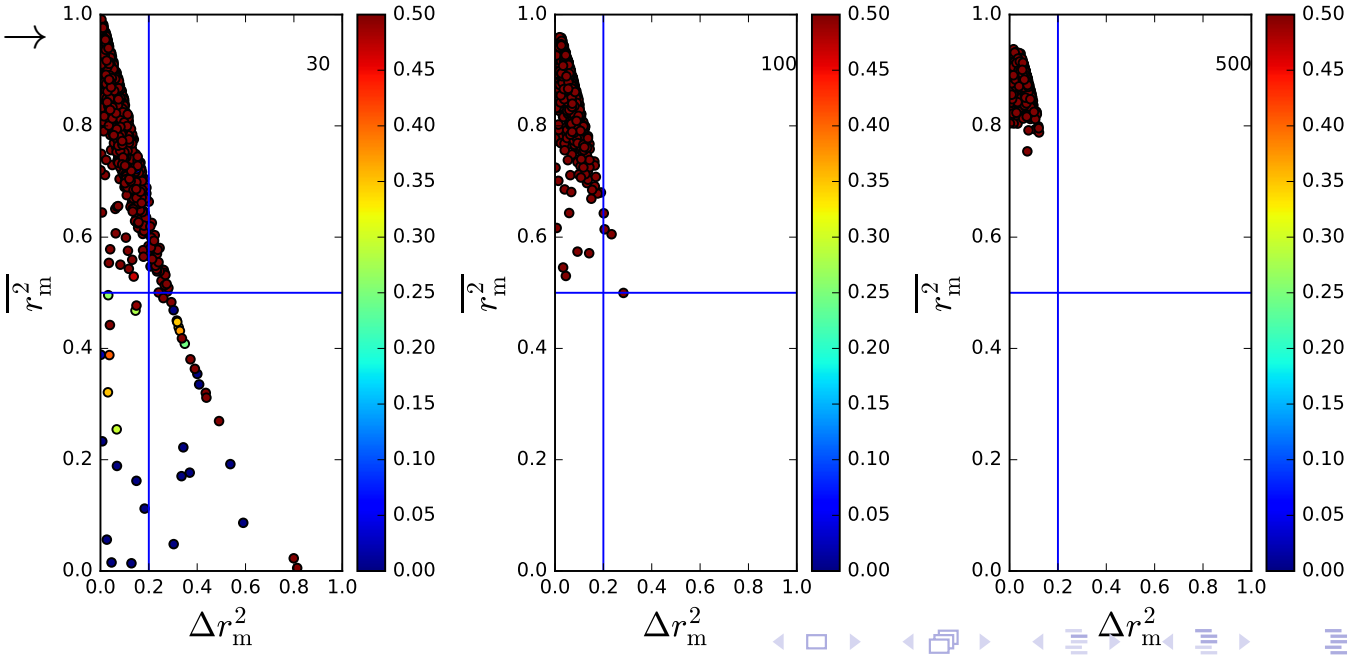
Q^2_{F2} a legérzékenyebb

Külső VPk - Roy-Ojha diagram

$R^2_{\text{beton}}=0.62\rightarrow$



$R^2_{\text{erőmű}}=0.92\rightarrow$

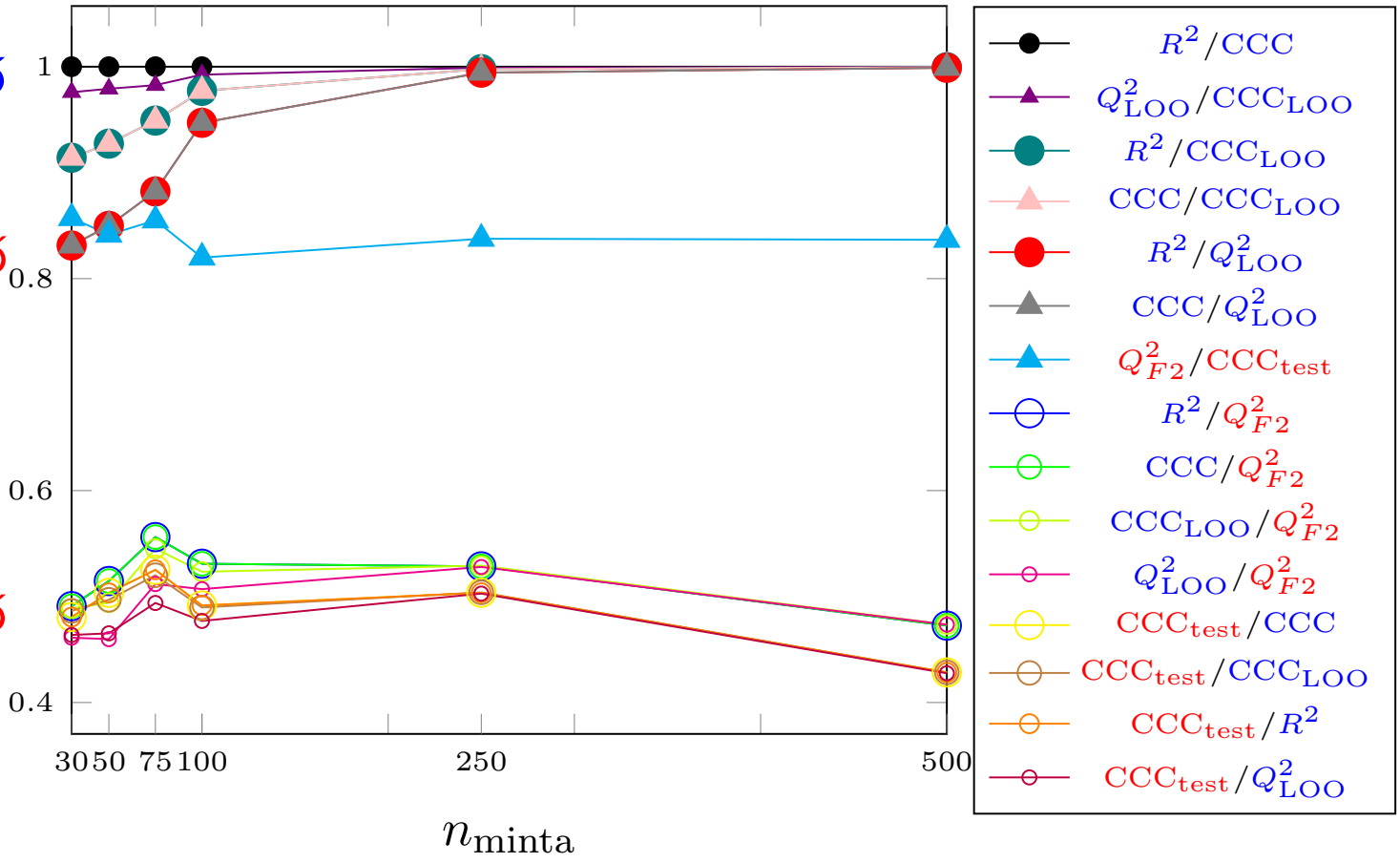


kiválasztott VPk: $\underbrace{R^2, \mathrm{CCC}, Q^2_{\mathrm{LOO}}, \mathrm{CCC}_{\mathrm{LOO}}}_{\text{belső}}, \underbrace{Q^2_{F2}, \mathrm{CCC}_{\mathrm{test}}}_{\text{külső}}$

belső/belső

külső/külső

belső/külső



Összegzés

- modell validáció: illesztés jósága, robusztusság & prediktivitás
- R^2, CCC, RMSE, MAE mintaméret függése anomáliát mutat
- külső validáció Q^2_{F1-3} paraméterei különböző mintaméretre külöböző érzékenységet mutatnak
- scrambling-gel kiválthatjuk a randomizációt
- Q^2_{LOO} használható az LMO változatok helyett
- 2D Roy-Ojha diagram kategorizálja a modellek prediktivitását
- minden mintaelemszámnál eltér a külső és belső validáció információtartalma

- nemlineáris modellezés, PLS

Munkatársak: Kovács Dániel, Tóth Gergely