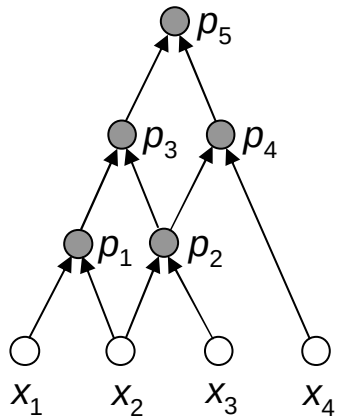
Rudjer Boskovic Institute
Division of electronics
Laboratory for information systems

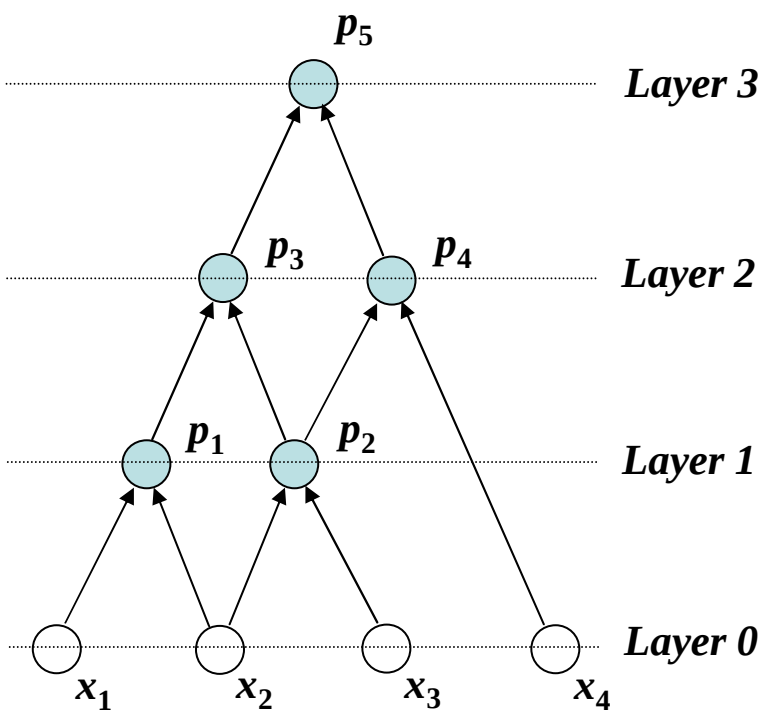
Ivan Marić

GMDH: building self-organizing feedforward perceptron-like polynomial models for real-time applications



GMDH: feedforward polynomial structure

GMDH example



Low-order polynomial

$$p = a_0 + a_1 z_i + a_2 z_j + a_3 z_i^2 + a_4 z_j^2 + a_5 z_i z_j$$

where z_i *and* z_j can be any variable from lower layers e.g.:

$$p_2 = p_2(x_2, x_3)$$
$$p_4 = p_4(p_2, x_4)$$

Real system: $y = f(x_1, x_2, x_3, x_4)$

GMDH approximation in recursive form:

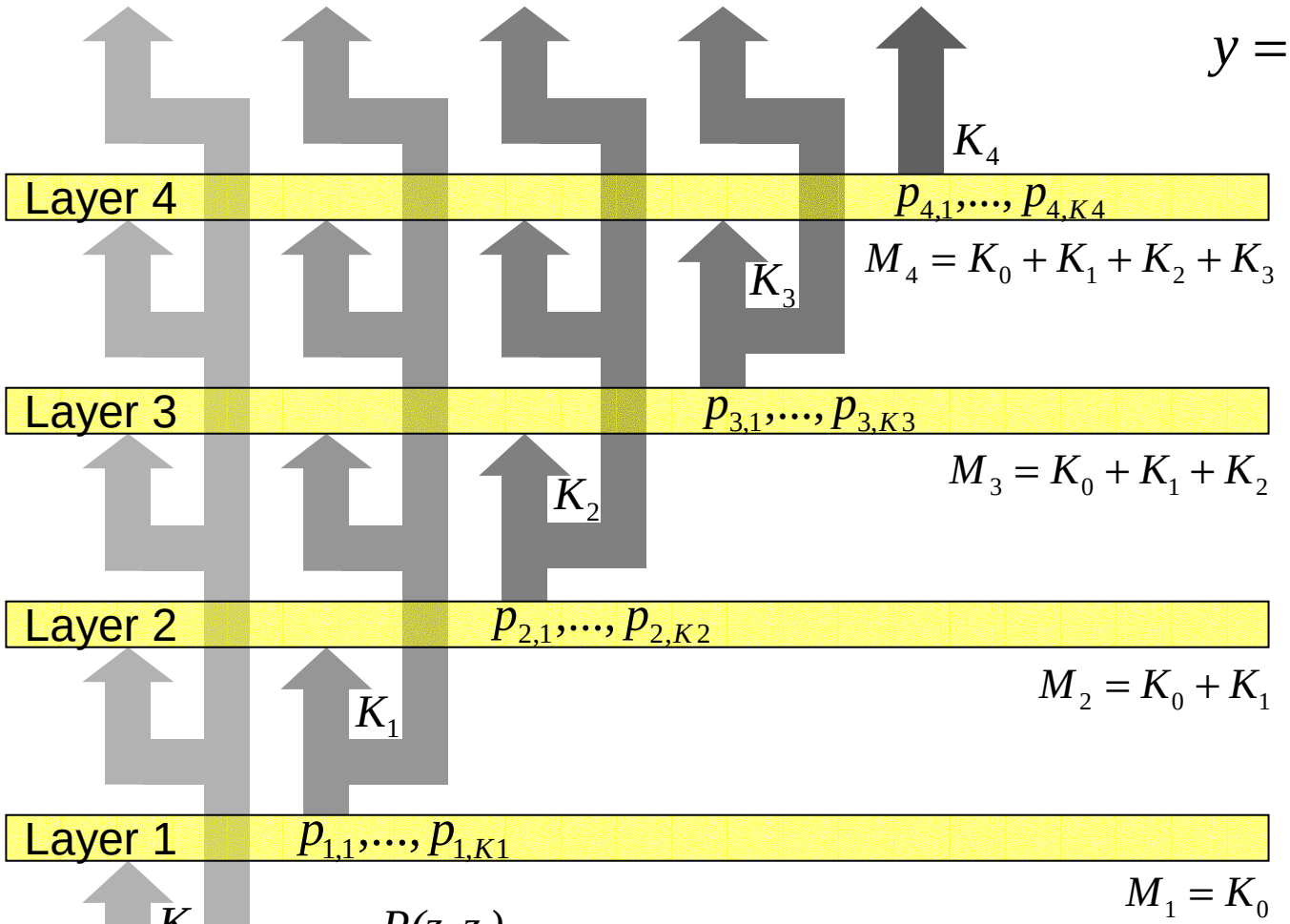
$$p_5 \approx y$$
$$p_5 = p_5(p_3(p_1(x_1, x_2), p_2(x_{2,} x_3)), p_4(p_2(x_{2,} x_3), x_4))$$

Kolmogorov-Gabor polynomial

$$P = a_0 + \sum_{i=1}^{N} a_i x_i + \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j x_i x_j + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} a_i a_j a_k x_i x_j xk + \ldots$$

GMDH algorithm (Polynomial theory of complex systems,IEEE Sys. Man Cyber., Ivakhnenko 1971)

$$y = y(x_1, x_2, \ldots, x_K$$

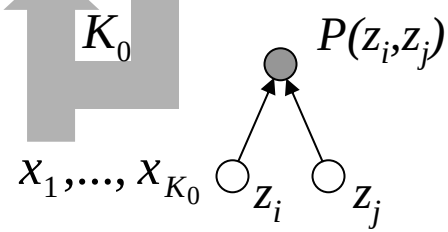


$$M_4 = K_0 + K_1 + K_2 + K_3 \qquad R_4 = \left(\frac{M_4 \cdot (M_4 - 1)}{2}\right)$$

$$M_3 = K_0 + K_1 + K_2 \qquad R_3 = \frac{M_3 \cdot (M_3 - 1)}{2}$$

$$M_2 = K_0 + K_1 \qquad R_2 = \frac{M_2 \cdot (M_2 - 1)}{2}$$

$$M_1 = K_0 \qquad R_1 = \frac{M_1 \cdot (M_1 - 1)}{2}$$



M_L – Total number of input variables to layer L
R_L – Total number of regression polynomials generated at layer L
K_L – Total number of new variables from layer L (adjustable)

Illustration of GMDH algorithm complexity: Total number of polynomials for 9 input variables

	No limitations on the total number of retained polynomials per layer		Maximum 50 polynomials retained per layer	
	Total number of input variables	Total number of possible polynomials	Total number of input variables	Total number of polynomials
Layer 1	9	36	9	36
Layer 2	45	990	45	990
Layer 3	1035	535095	95	4465
Layer 4	536130	143717420385	145	10440
Layer 5	143717956515	1.03274255123e+22	195	18915
Layer 6	1.03274255124e+22	5.33278588580e+43	245	29890
Layer 7	5.33278588580e+43	1.42193026519e+87	295	43365

Polynomial regression

Learning data set used for polynomial regression

$$\left\{x_{1i}^{L}, x_{2i}^{L}, \ldots, x_{Ki}^{L}, y_{i}^{L}\right\}; i = 1, \ldots, M$$

Low order polynomial

$$p = a_0 + a_1 z_i + a_2 z_j + a_3 z_i^2 + a_4 z_j^2 + a_5 z_i z_j$$

Set of 6 simultaneous linear equations

$$\frac{\partial}{\partial a_k}\left(\sum_{m=1}^{M}\left(y_m^L - a_0 - a_1 z_{i,m}^L - a_2 z_{j,m}^L - a_3 z_{i,m}^{L\ 2} - a_4 z_{j,m}^{L\ 2} - a_5 z_{i,m}^L z_{j,m}^L\right)^2\right) = 0$$

$$k = 0, \ldots, 5$$

Set of 6 simultaneous linear equations

$$\sum_{m=1}^{M}\left(a_0 + a_1 z_{i,m}^{L} + a_2 z_{j,m}^{L} + a_3 z_{i,m}^{L\ 2} + a_4 z_{j,m}^{L\ 2} + a_5 z_{i,m}^{L} z_{j,m}^{L} - y_m\right) = 0$$

$$\sum_{m=1}^{M} z_{i,m}^{L}\left(a_0 + a_1 z_{i,m}^{L} + a_2 z_{j,m}^{L} + a_3 z_{i,m}^{L\ 2} + a_4 z_{j,m}^{L\ 2} + a_5 z_{i,m}^{L} z_{j,m}^{L} - y_m\right) = 0$$

$$\sum_{m=1}^{M} z_{j,m}^{L}\left(a_0 + a_1 z_{i,m}^{L} + a_2 z_{j,m}^{L} + a_3 z_{i,m}^{L\ 2} + a_4 z_{j,m}^{L\ 2} + a_5 z_{i,m}^{L} z_{j,m}^{L} - y_m\right) = 0$$

$$\sum_{m=1}^{M} z_{i,m}^{L\ 2}\left(a_0 + a_1 z_{i,m}^{L} + a_2 z_{j,m}^{L} + a_3 z_{i,m}^{L\ 2} + a_4 z_{j,m}^{L\ 2} + a_5 z_{i,m}^{L} z_{j,m}^{L} - y_m\right) = 0$$

$$\sum_{m=1}^{M} z_{j,m}^{L\ 2}\left(a_0 + a_1 z_{i,m}^{L} + a_2 z_{j,m}^{L} + a_3 z_{i,m}^{L\ 2} + a_4 v + a_5 z_{i,m}^{L} z_{j,m}^{L} - y_m\right) = 0$$

$$\sum_{m=1}^{M} z_{i,m}^{L} z_{j,m}^{L}\left(a_0 + a_1 z_{i,m}^{L} + a_2 z_{j,m}^{L} + a_3 z_{i,m}^{L\ 2} + a_4 z_{j,m}^{L\ 2} + a_5 z_{i,m}^{L} z_{j,m}^{L} - y_m\right) = 0$$

Best model selection

Test data are used for the selection of the best models

$$\left\{x_{1i}^T, x_{2i}^T, \ldots, x_{Ki}^T, y_i^T\right\};\ j = 1, \ldots, N$$

<u>BEST MODEL SELECTION CRITERIA</u>

- $\boldsymbol{E_{ls}}$: Least Square Error Measure

$$E_{ls} = \sum_{i=1}^{N} \left(p_i^T - y_i^T\right)^2$$

- $\boldsymbol{E_{rrs}}$: Root Relative Squared Error Measure

$$E_{rrs} = \sqrt{\frac{\sum_{i=1}^{N} (p_i^T - y_i^T)^2}{\sum_{i=1}^{N} (y_i^T - \bar{y})^2}}$$

- $\boldsymbol{E_{CE}}$: Compound Squared Relative Error Measure

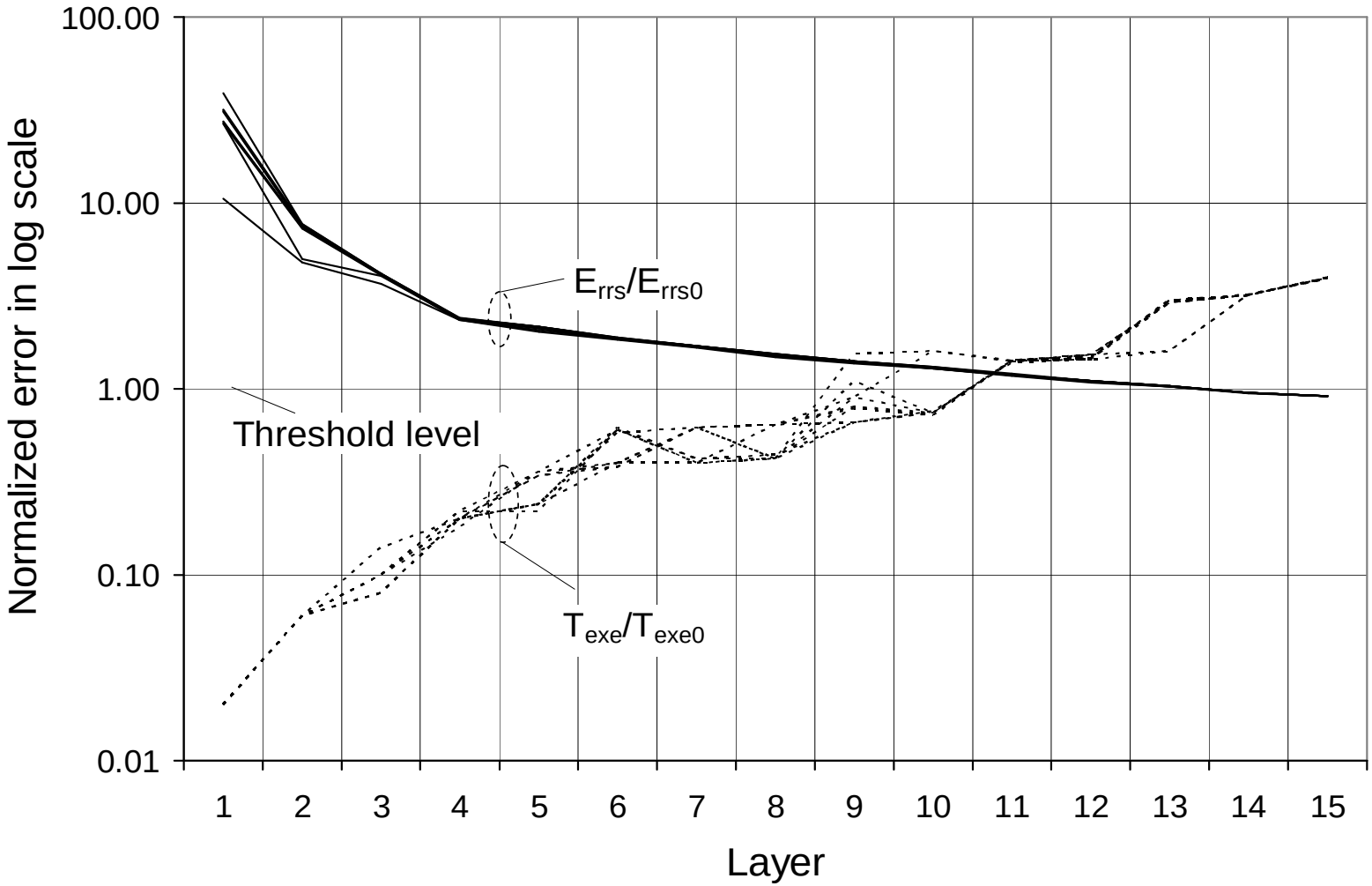
$$E_{CE} = c_w \left(\frac{E_{rrs}}{E_{rrs0}}\right)^2 + \left(1 - c_w\right)\left(\frac{T_{exe}}{T_{exe0}}\right)^2$$

T_{exe} – model execution time

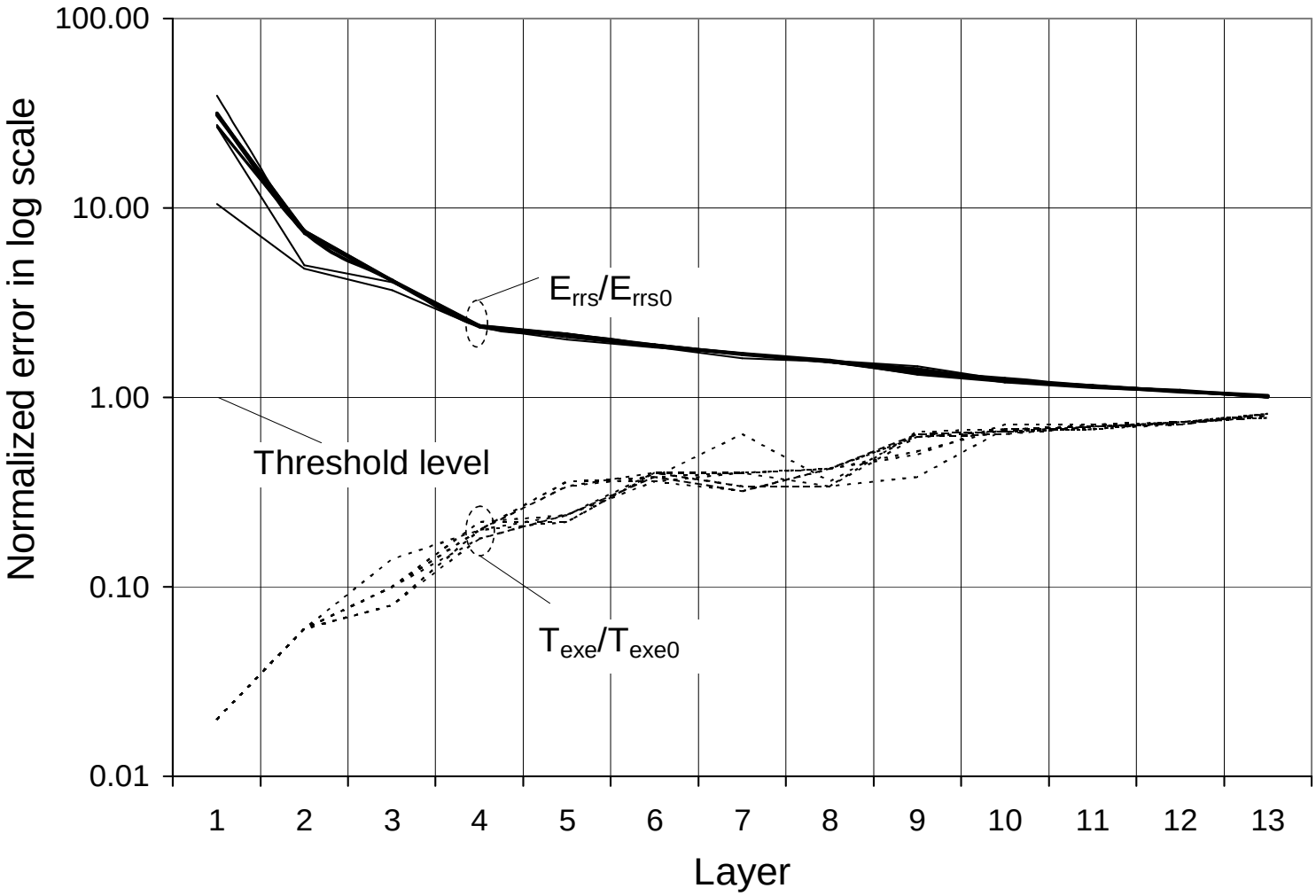
$0 \le c_w \le 1$ – weighting coefficient

E_{rrs0}, T_{exe0} – thresholds

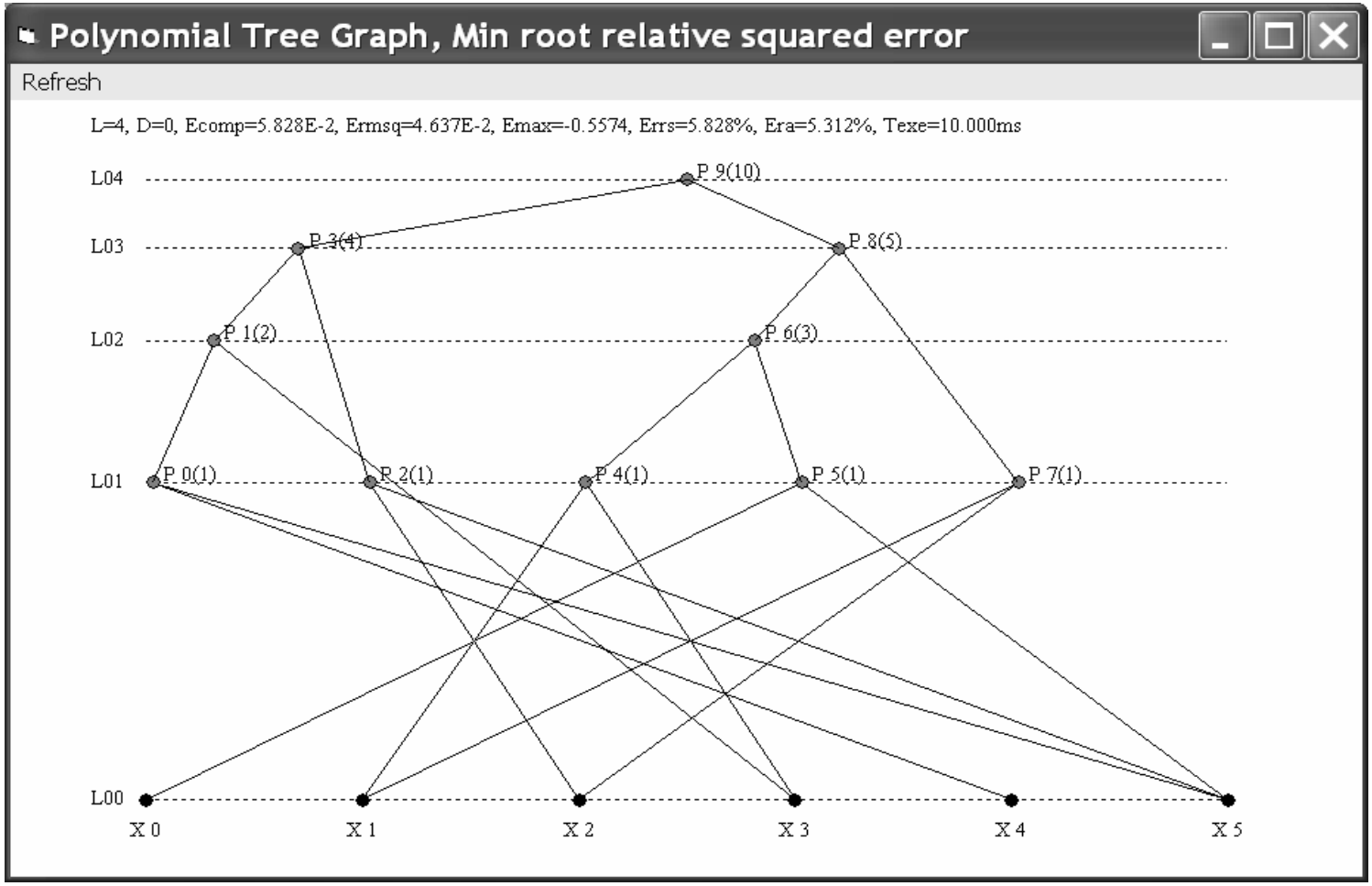
The approximation error and the execution time when using the LS or RRS error criterion for model selection



The approximation error and the execution time when using the compound error criterion for model selection



An example of a GMDH polynomial model obtained when using RRS error mesure for the selection of the best candidate models



y=P9(P3(P1(P0(x4,x5),x3),P2(x2,x5)),P8(P6(P4(x1,x3),P5(x0,x5)),P7(x1,x2)))

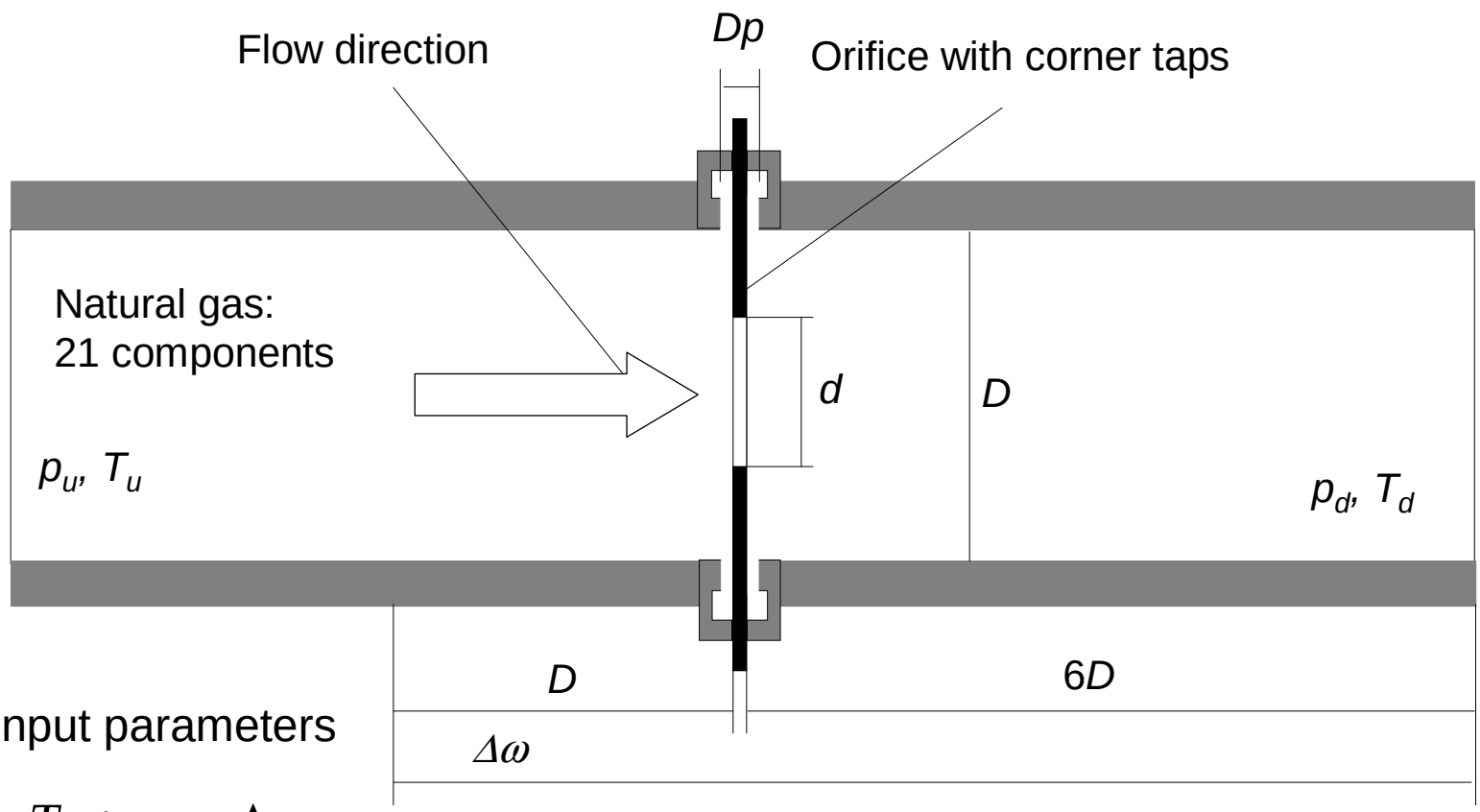
GMDH polynomial equation in recursive form

$$P = P_9\left(P_3\left(P_1\left(P_0\left(x_4, x_5\right), x_3\right), P_2\left(x_2, x_5\right)\right), P_8\left(P_6\left(P_4\left(x_1, x_3\right), P_5\left(x_0, x_5\right)\right), P_7\left(x_1, x_2\right)\right)\right)$$

$$P_i = a_0 + a_1 x_j + a_2 x_k + a_3 x_j^2 + a_4 x_k^2 + a_5 x_j x_k$$

i	a0	a1	a2	a3	a4	a5
0	-1.3735E+01	1.1914E+01	5.4858E-01	5.4678E+00	-2.9526E-03	-3.5912E-01
1	8.1678E+00	2.8764E+00	-5.2126E-02	1.9260E-02	8.6716E-05	-6.8152E-03
2	2.4070E+00	-1.0205E-01	1.2623E-01	-7.9747E-03	-1.8498E-03	1.5925E-03
3	3.7246E+00	-8.7866E-01	-1.0388E+00	4.3489E-03	2.5581E-02	4.7307E-01
4	1.9692E+01	-4.6700E+00	-7.8230E-02	-4.2119E+01	8.5175E-05	1.8046E-02
5	9.7814E+00	-1.7134E+01	-4.0561E-01	1.2834E+01	6.4154E-03	6.4112E-01
6	2.2563E+00	-2.4485E-01	-9.0911E-01	8.2653E-04	8.5049E-02	3.1820E-01
7	4.5270E+00	1.8325E+00	-4.0275E-02	-3.8016E+01	-7.9662E-03	-1.0699E-01
8	3.2241E+00	-7.2378E-01	-9.2336E-01	-7.5029E-03	1.8362E-02	4.5607E-01
9	7.9213E-02	1.0500E-01	8.4703E-01	1.6364E-01	7.6373E-02	-2.3335E-01

Compensation of natural gas flow-rate error due to temperature drop effect



26 input parameters

$$T_u = T_d + \mu_{JT} \cdot \Delta\omega$$

$$q_u = q(P_u, T_u, \Delta p, \rho_u, \gamma_u, \kappa_u, D, d)$$

$$q_d = q(P_u, T_d, \Delta p, \rho_d, \gamma_d, \kappa_d, D, d)$$

$$K_q = q_u / q_d \qquad q_{GMDH} = K_{GMDH} \cdot q_d$$

$$K_{GMDH} \approx K_q \qquad E_{GMDH} = (q_{GMDH} - q_u)/q_u$$

Flow-rate correction factor modeling

Input parameters (26): $x_1, x_{2,} ..., x_{21}, p_u, T_d, \Delta p, D, d$

PREPROCESSING

- Random generation of learning set (20000 samples)
- Random generation of test set (20000 samples)
- Reduction of input parameters (exp. knowledge): 26 to 13

$$x_{CO_2}, x_{H2}, p_u, T_d, \rho_{r_d}, H_S, \rho_d, \gamma, \kappa, \Delta p, D, d, \beta$$

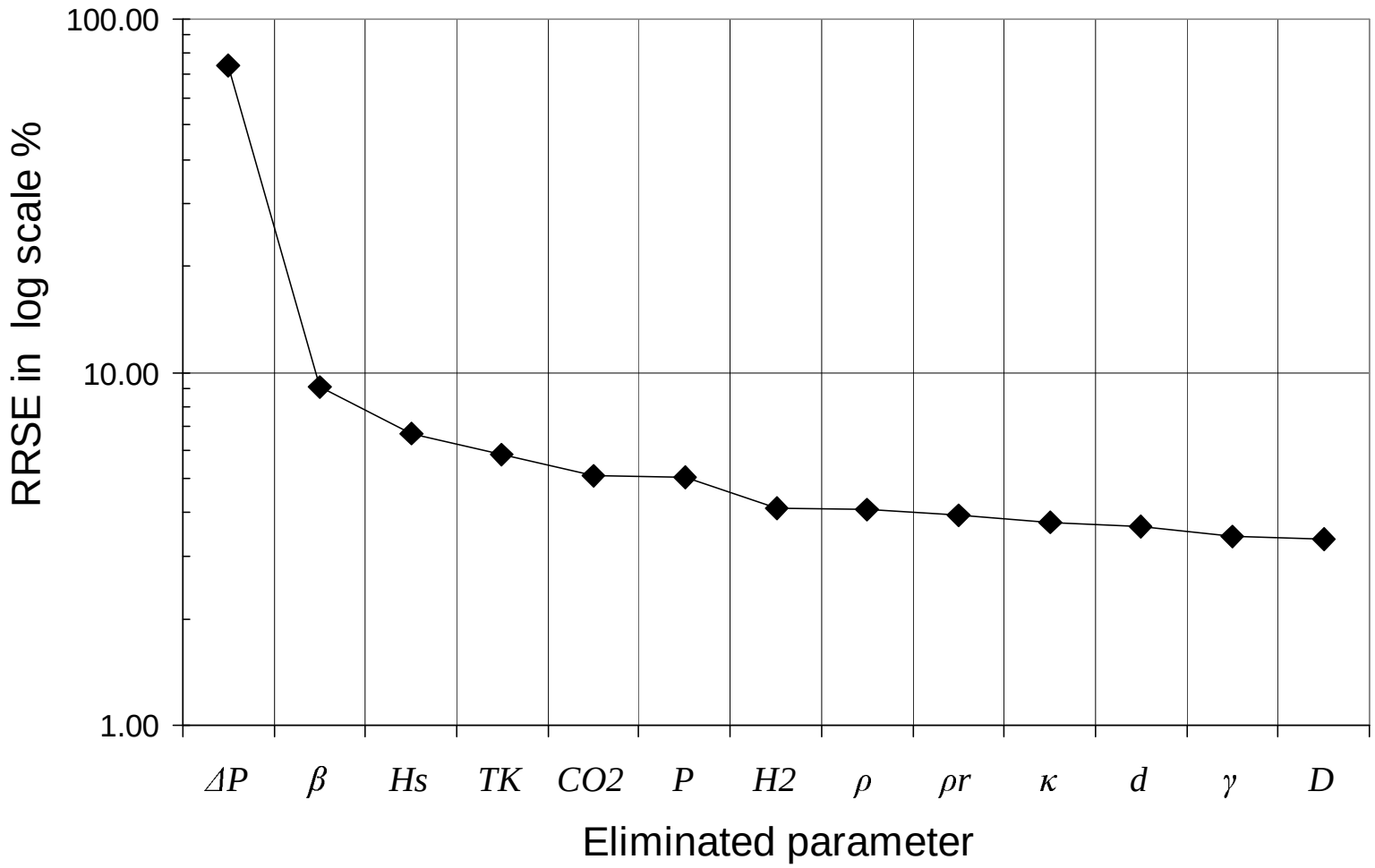
- Sorting the parameters in order of significance
- Determination of optimal input parameters (9 out of 13):

$$x_{CO_2}, x_{H2}, p_u, T_d, \Delta p, \rho_d, \rho_{r_d}, H_S, \beta$$

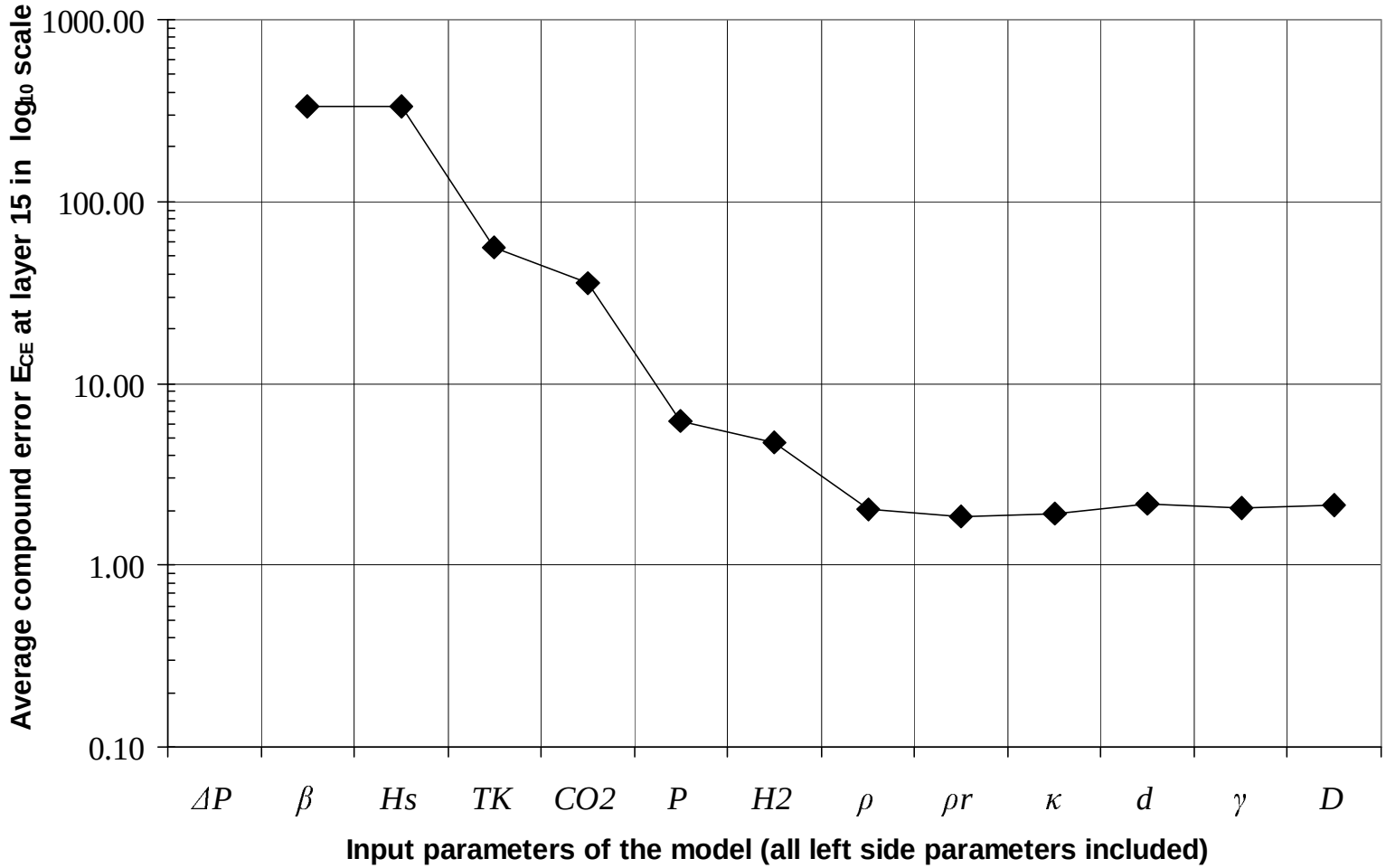
MODELING

- Maximum execution time of low order polynomial: 1ms
- Execution time of the flow-rate correction: ≤50ms
- RRSE of the flow-rate correction factor: ≤4%
- Total number of layers: ≤15
- Total number of retained polynomials per layer: ≤25, ≤50, ≤75, ≤100
- Model selection criterion: CE (c_w=0, 0.1, 0.2,..., 1.0)

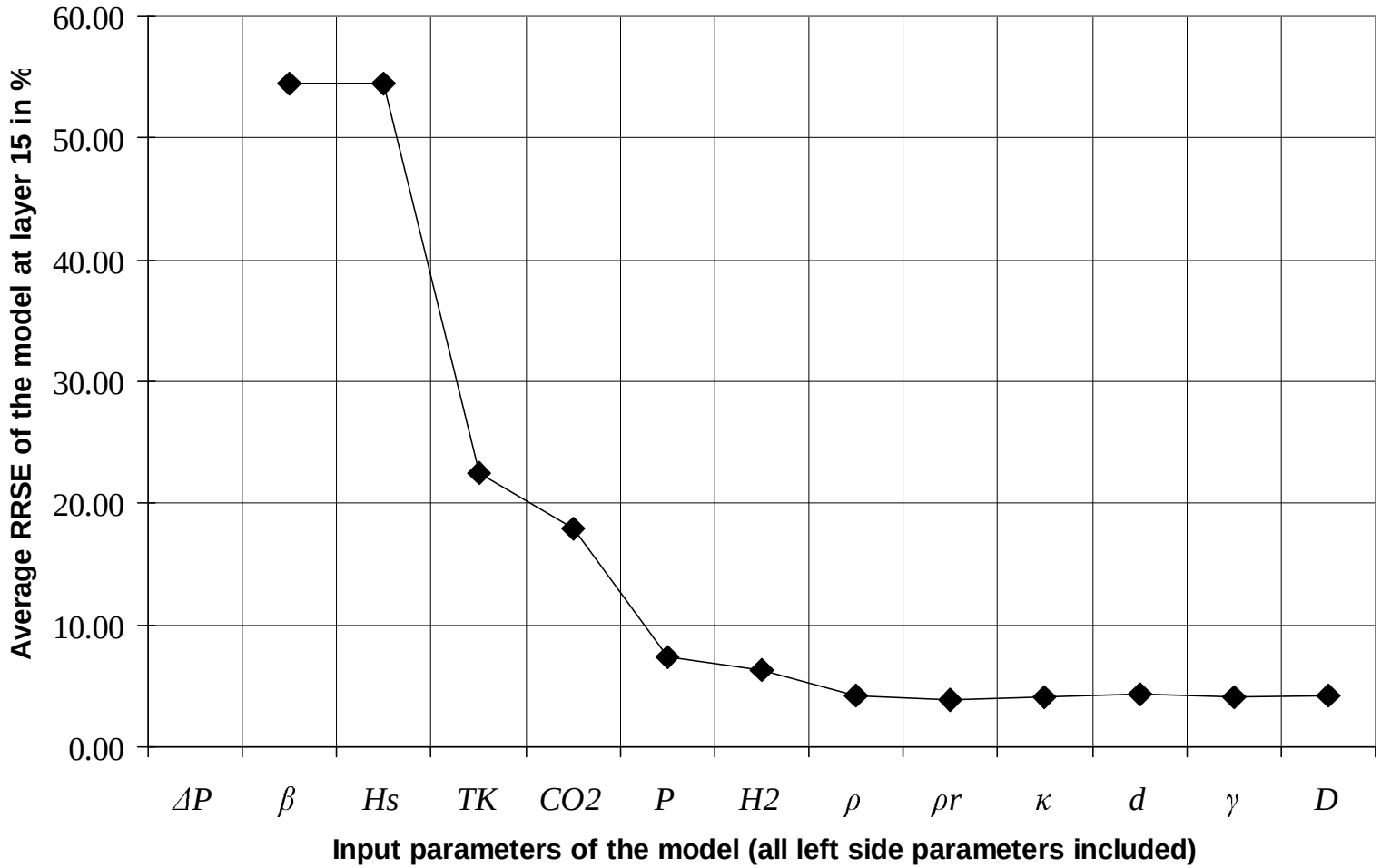
Average RRSE of the best 10 polynomial models from each layer related to the single input variable, as the result of its exclusion from the set



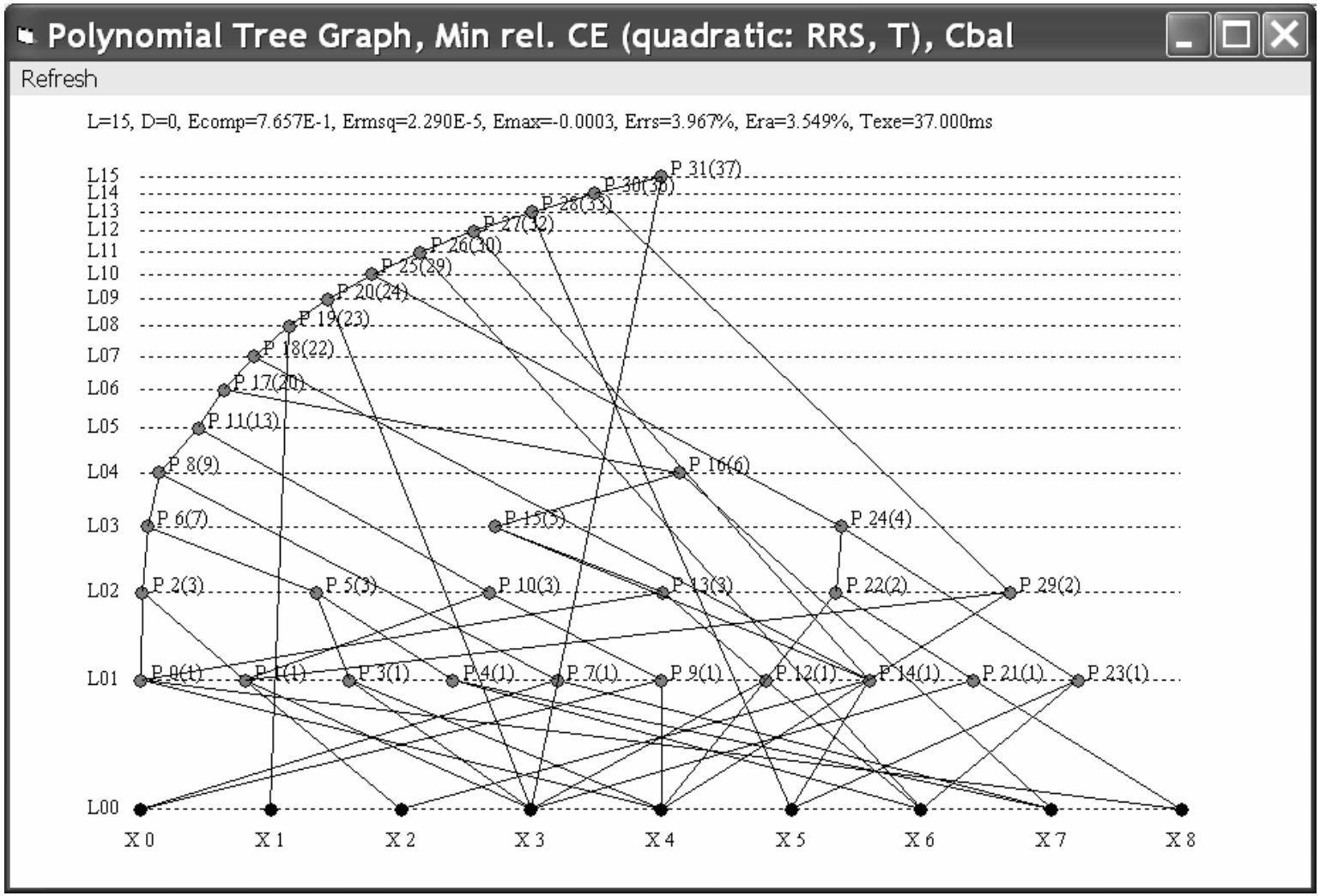
Average CE for the 10 best models obtained at layer 15



Average RRSE for the 10 best models obtained at layer 15



The best GMDH surrogate of the flow-rate correction factor *Kq*, satisfying the prespecified conditions, is obtained at layer 15 by using the CE measure (c_w=0.5, RRSE=3.967%, ET=37ms)



The best GMDH surrogate model of the flow-rate correction factor, obtained at layer 15 by using the CE measure (c_w=1.0, RRSE=3.915%, ET=197ms), fails to satisfy the prespecified conditions

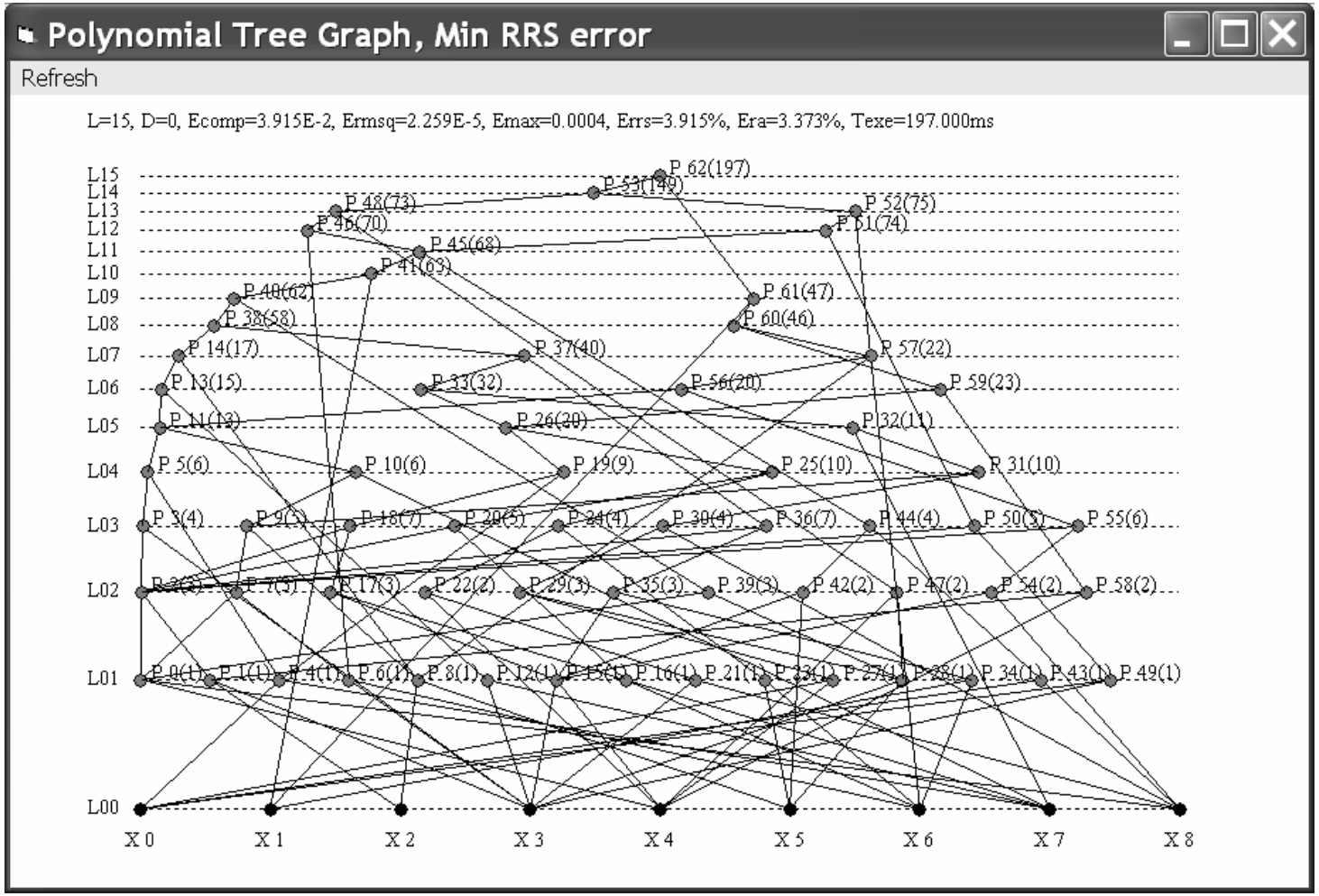


Illustration of relative error in the measurement of a natural gas flow-rate by orifice plates with corner taps when ignoring the temperature drop effect (no correction)

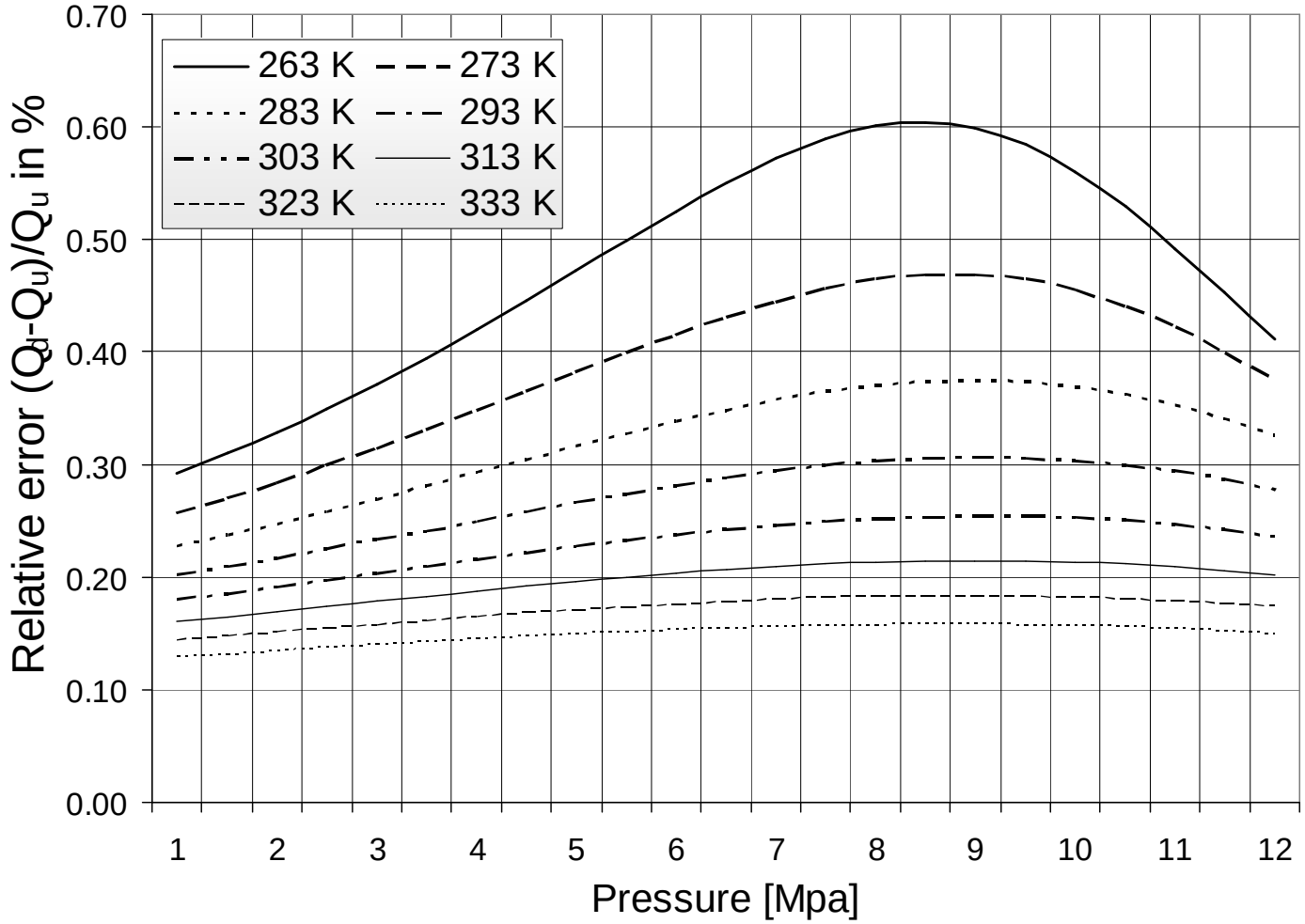
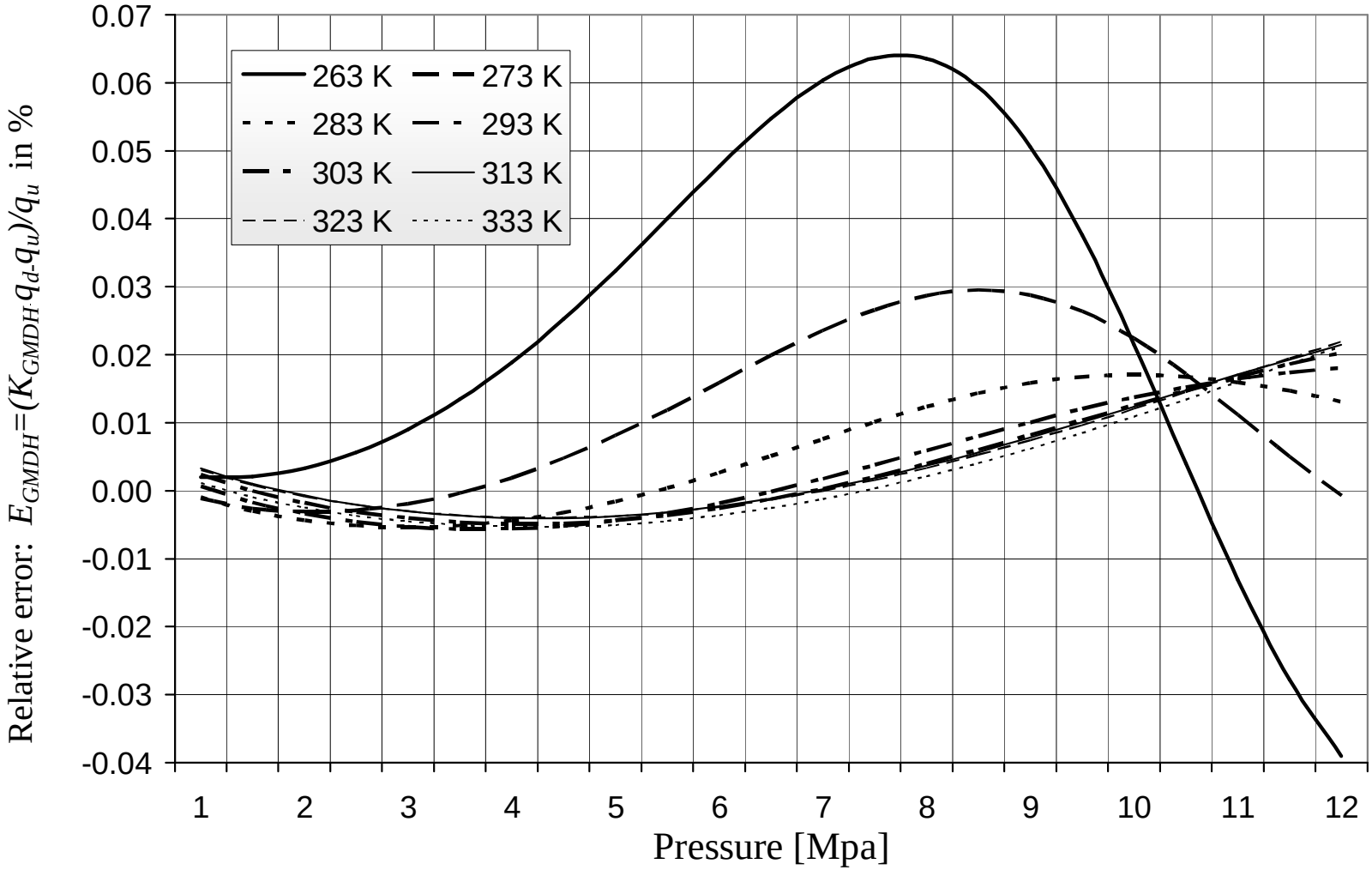


Illustration of relative error in the measurement of a natural gas flow-rate, when using the GMDH polynomial for the correction of the temperature drop effect



Modeling natural gas properties

- Input parameters (6):

$$p, T, H_S, \rho, x_{CO2}, x_{H2}$$

- PREPROCESSING
- Random generation of learning set (20000 samples)
- Random generation of test set (20000 samples)

- MODELING
- Maximum execution time of low order polynomial: 1ms
- Execution time of the flow-rate correction: ≤50ms
- RRSE of the flow-rate correction factor: ≤3%
- Total number of layers: ≤15
- Total number of retained polynomials per layer: ≤25, ≤50, ≤75, ≤100
- Model selection criterion: CE (*cw*=0, 0.1, 0.2,..., 1.0)

The best GMDH model of JT coefficient, satisfying the prespecified conditions, is obtained at layer 13 by using the CE measure for model selection (c_w=0.5, RRSE=2.493%, ET=41ms)

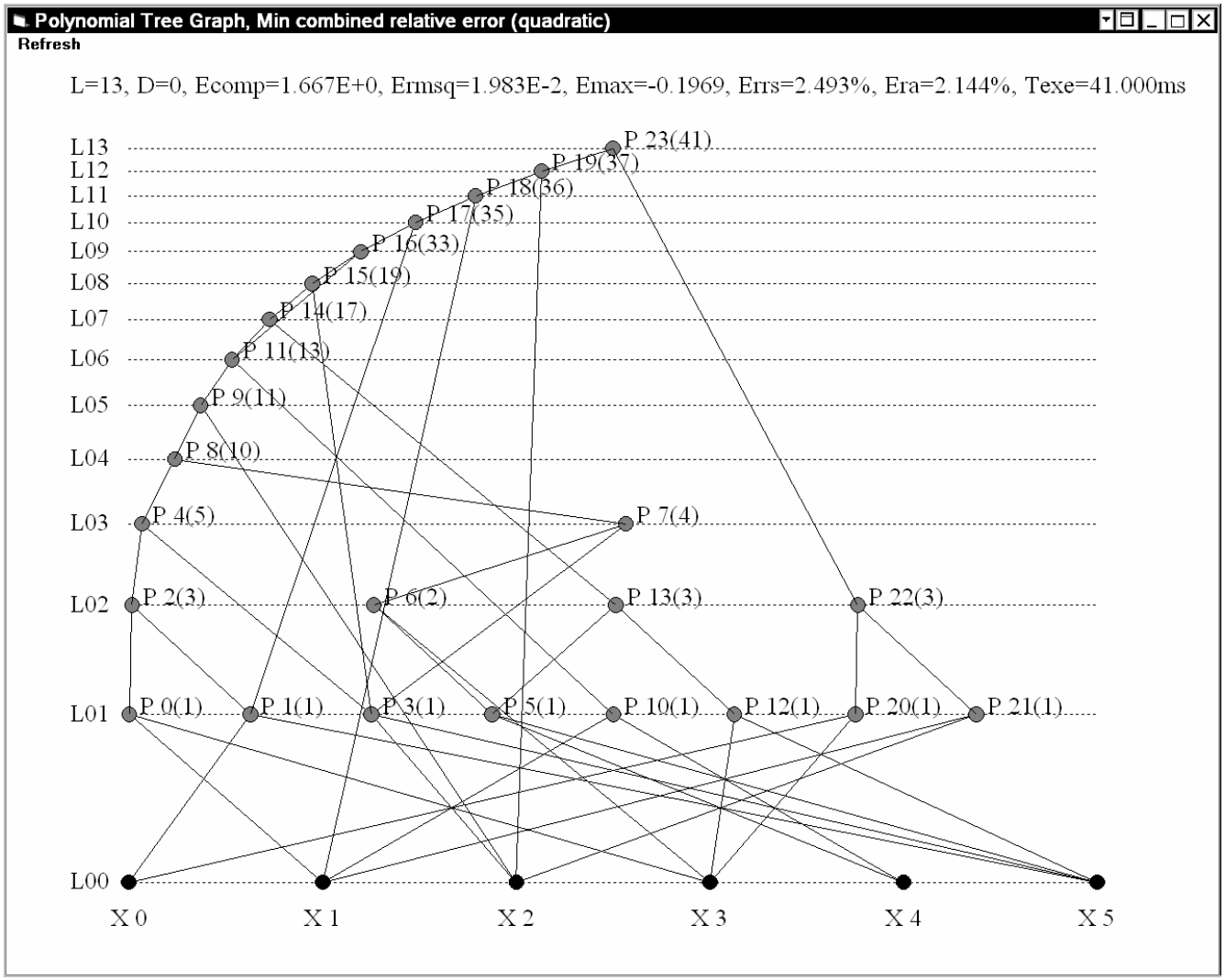
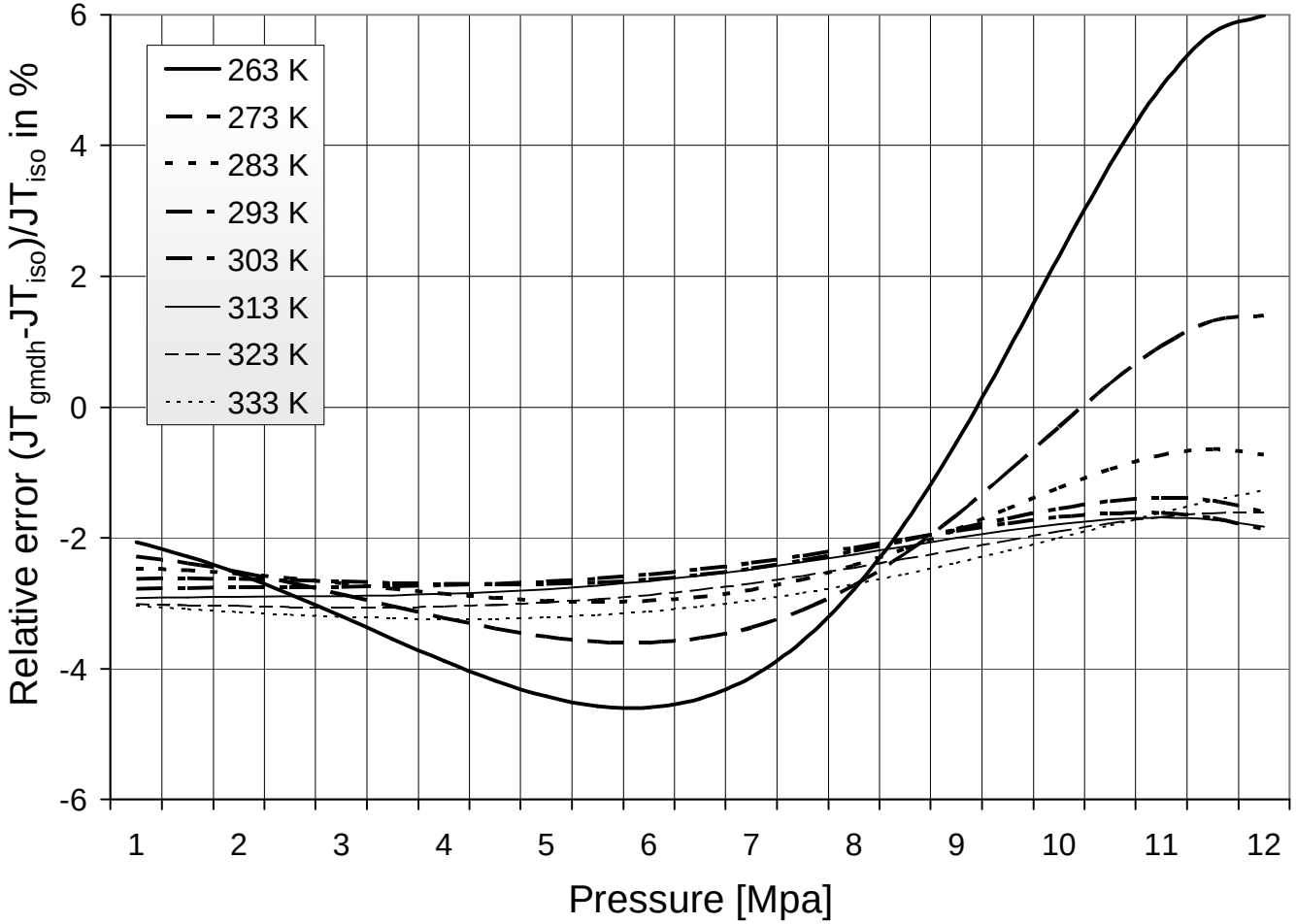


Illustration of relative error of JT coefficient due to its approximation by GMDH model



The best GMDH model of molar heat capacity, satisfying the prespecified conditions, is obtained at layer 13 by using the CE measure for model selection (c_w=0.5, RRSE=2.995%, ET=47ms)

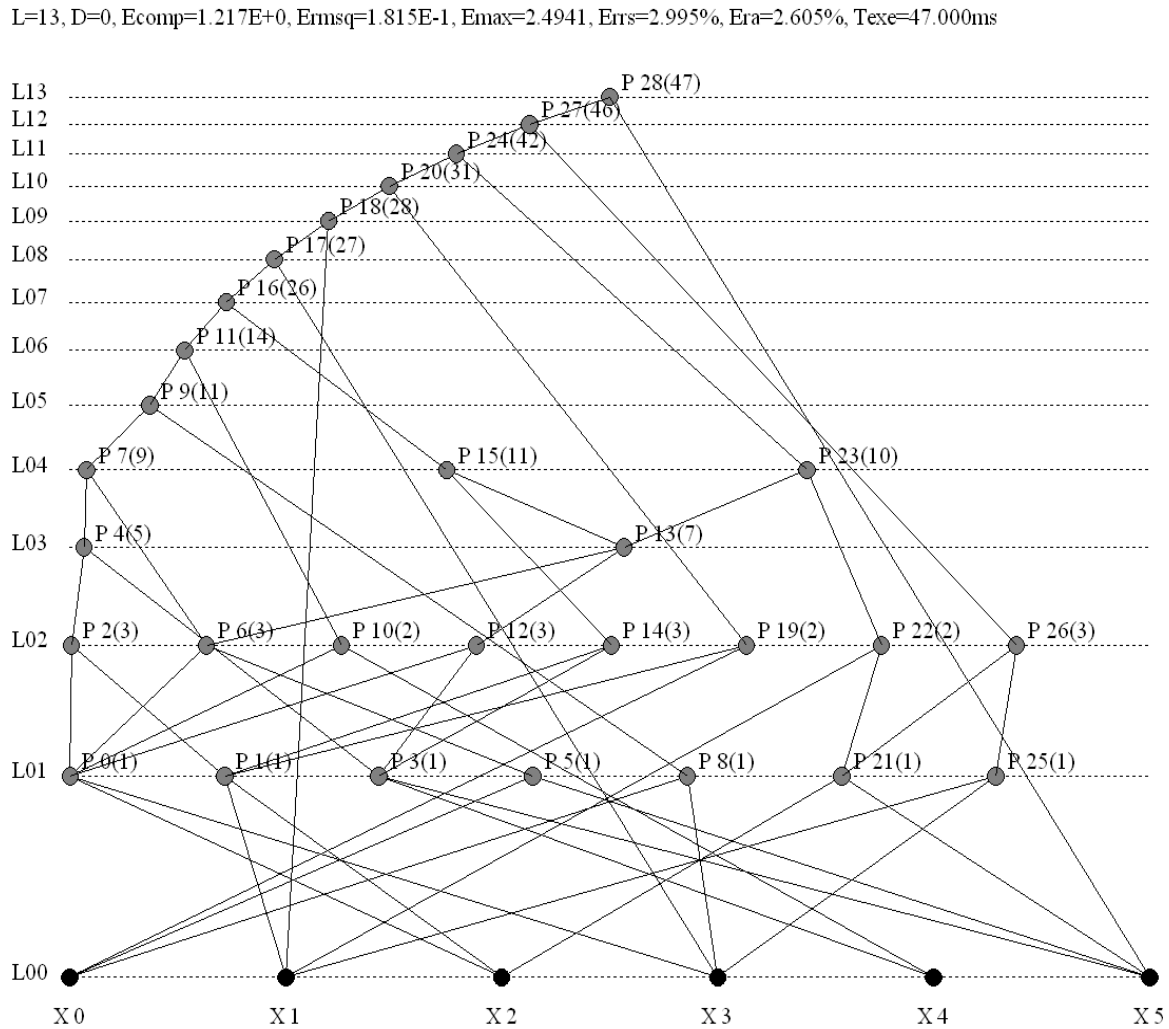
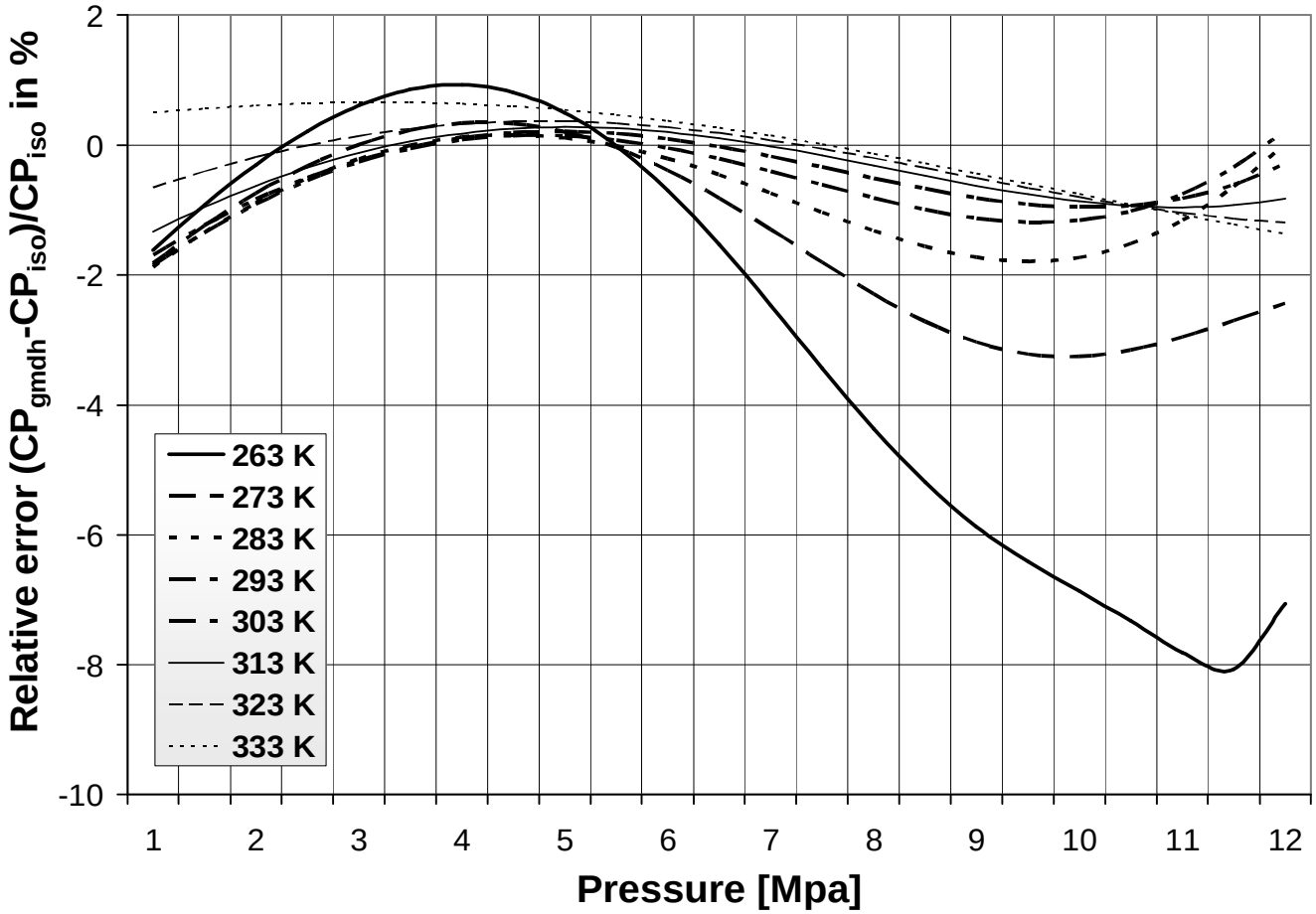


Illustration of relative error of molar heat capacity due to its approximation by GMDH model



The best GMDH model of isentropic exponent, satisfying the prespecified conditions, is obtained at layer 14 by using the CE measure for model selection (c_w=0.5, RRSE=2.996%, ET=33ms)

L=14, D=0, Ecomp=1.433E+0, Ermsq=2.861E-3, Emax=0.0368, Errs=2.996%, Era=2.631%, Texe=33.000ms

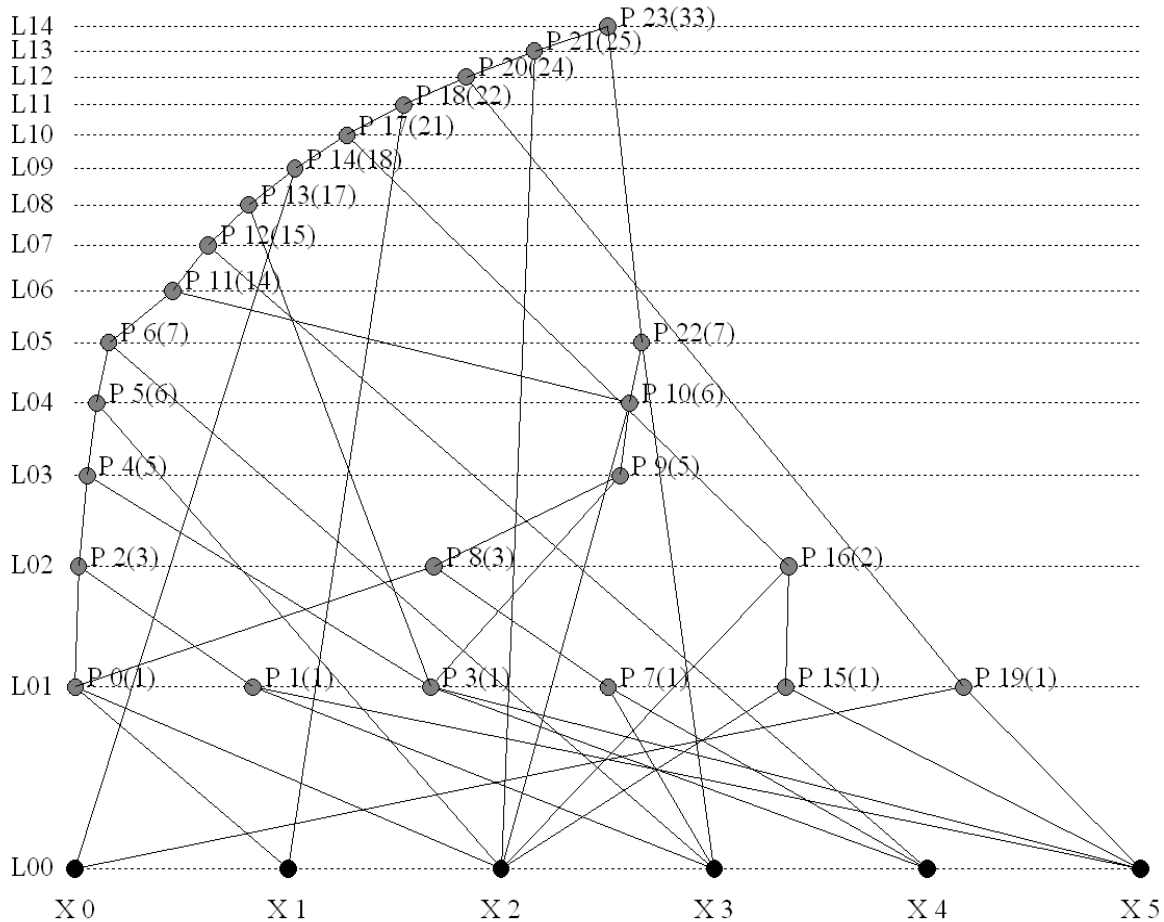
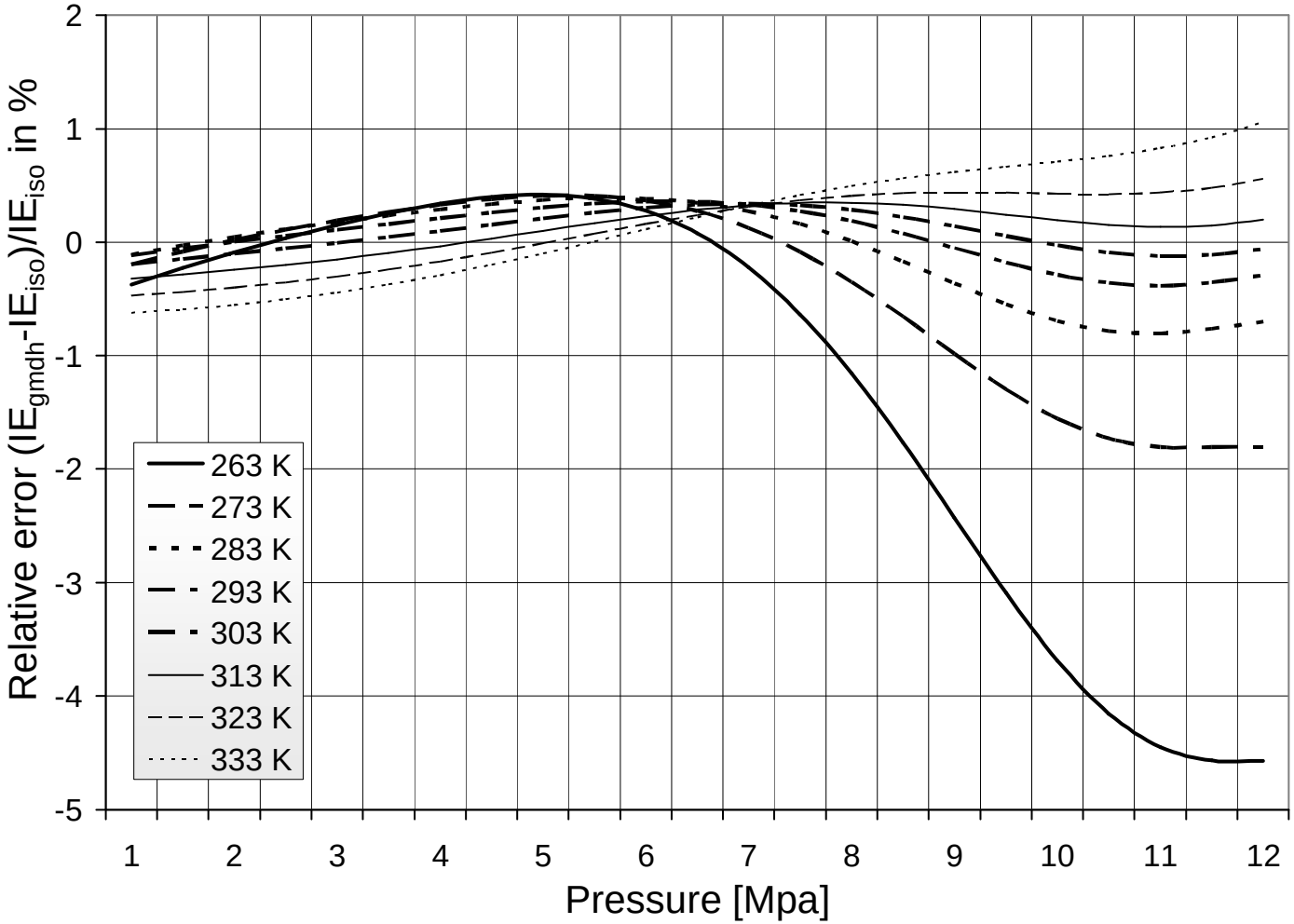
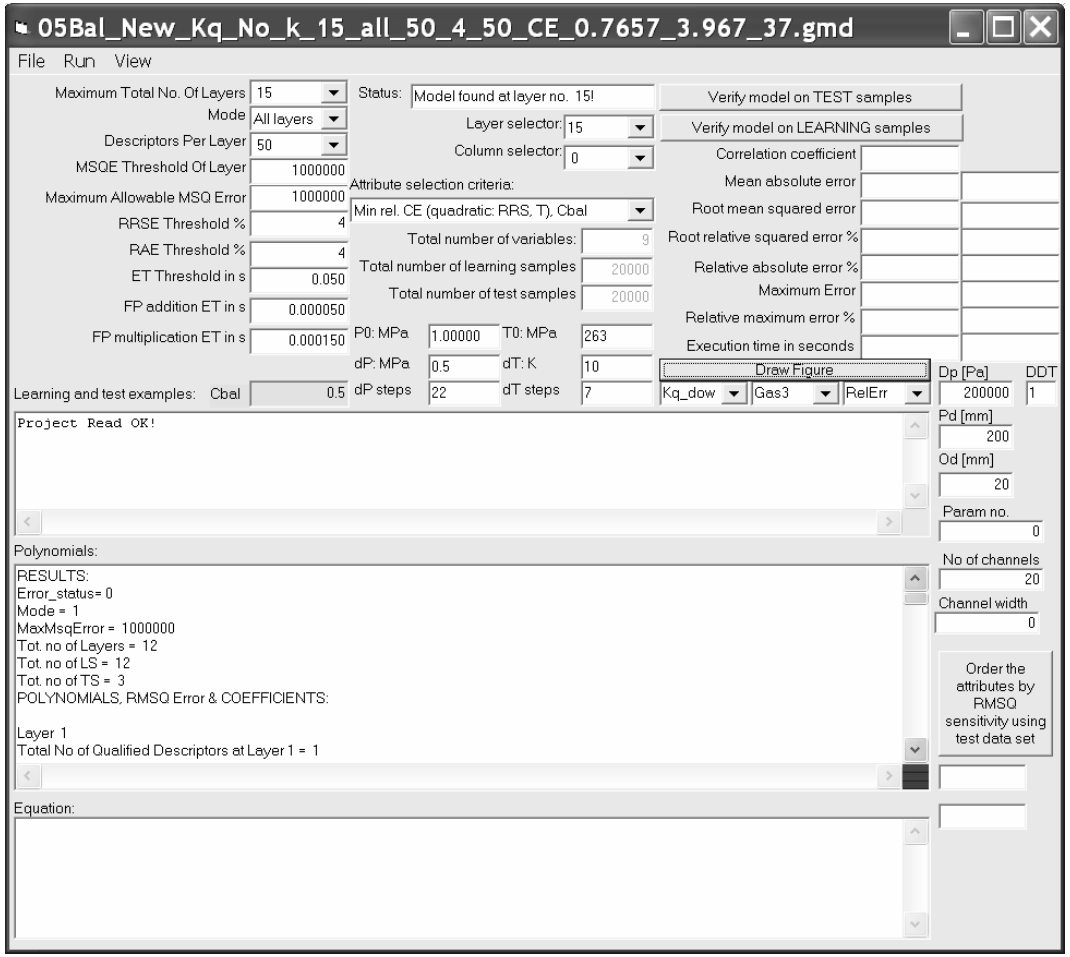


Illustration of relative error of the isentropic exponent due to its approximation by GMDH model



User interface to our GMDH system



Conclusions

- CE measure proves to be very efficient when building the GMDH models for real-time application
- It forces the GMDH algorithm to tailor the model with respect to the accuracy and the complexity
- It makes complex procedures feasible in real-time with acceptable degradation of approximation accuracy
- It can be modified to enable model generation by controling multiple parameters
- GMDH algorithm can be applied to modeling the complex problems in science and in economy