Construction and Applications of Significant Polyhedra

Klaus Truemper

Department of Computer Science
University of Texas at Dallas
Richardson, TX 75083
U.S.A.

Definitions

$\mathcal{E}$= some process

x = vector in R^n

t = scalar

$X = \{(x,t)$ instances$\}$
= sample of data collected from $\mathcal{E}$

I = interval of t

P = polyhedron in R^n

P is always full-dimensional, and some defining inequalities may be strict.

Problem

Find all intervals I and polyhedra P such that

1. The definition of P is comprehensible by humans in terms of process $\mathcal{E}$.

2. $\forall (x,t) \in X\colon t \notin I \Rightarrow x \notin P$.

3. With high probability, the *subgroup*
   $S = \{x \mid (x,t) \in X; x \in P\}$
corresponds to an unusual aspect of process $\mathcal{E}$. P and S are said to be *significant* for process $\mathcal{E}$.

Logic Formula

View P as a propositional logic formula $R(x)$ that is a conjunction whose literals are inequalities $a^t x \leq b$ or $a^t x < b$.

Example: $R(x) = (x_1 < 6.5) \wedge (x_1 + x_2 > 7.5) \wedge (x_1 - x_2 < 4.5)$

Subgroup Discovery Problem

As before: $X = \{(x, t)\}$ is a sample of a process $\mathcal{E}$.

Scalar t is a *target*.

Find all target intervals I and rules $R(x)$ such that

1. Humans can comprehend $R(x)$ in terms of process $\mathcal{E}$.

2. $\forall (x, t) \in X$: $t \notin I \Rightarrow R(x) = \mathit{False}$

3. With high probability, the *subgroup*
   $S = \{x \mid (x, t) \in X; R(x) = \mathit{True}\}$

corresponds to an unusual aspect of process $\mathcal{E}$.
$R(x)$ and S are said to be *significant* for $\mathcal{E}$.

Related Facts and Results

1. If there are essentially identical I and $R(x)$ cases, selection of a representative is acceptable.

2. A possible conclusion is that no significant rules exist about X.

Size and Comprehensibility of Formulas

Human comprehension of data or statements is an extensively covered topic of Neurology and Psychology.

Chunk: Collection of concepts that are closely related and have much weaker connections with other concurrently used concepts.

G. A. Miller (1956): "Magical number seven, plus or minus two" of chunks is limit of short-term memory storage capacity.
(10,851 citations)

N. Cowan (2001): "Magical number 4 of chunks.

G. S. Halford and N. Cowan (2005): Integrated treatment of *working memory capacity* and *relational capacity*.

(1) Working memory is limited to approximately 3-4 chunks.
(2) Number of variables involved in reasoning is limited to 4.

Implications for Subgroup Discovery

1. Human comprehension requires the inequalities to have at most 4 (1?, 2?, 3?) coefficients. Hence will consider only such formulas. Human processing of such an inequality amounts to *elementary chunking*.

2. Using Halford and Cowan (2005) and a reasonable assumption, formulas are comprehensible by humans if they have at most 4 (3?) literals.

Restated Subgroup Discovery Problem

Find all target intervals I and conjunctions $R(x)$ with linear inequalities as terms such that

1. There are at most 4 inequalities in $R(x)$, each of which has at most 4 nonzero coefficients.

2. $\forall (x,t) \in X\colon t \notin I \Rightarrow R(x) = \mathit{False}$

3. With high probability, the subgroup
   $S = \{x \mid (x,t) \in X; R(x) = \mathit{True}\}$

corresponds to an unusual aspect of process $\mathcal{E}$. $R(x)$ and S are said to be *significant*.

Some Complications

1. The dimension n of the vectors x may be large relative to the number N of vectors in X.

Example: $n = 100$ and $N = 30$.

2. Subvectors of x vectors may depict functions. For example, x_1, $x_2, \ldots,$ x_k may be measurements of one variable at k time points. This case always arises when longitudinal study data are processed.

Thus, the subgroup must represent functions. Can be done by computing characteristics of functions and constructing rules that use these characteristics.

Uses of Subgroup Discovery

1. Expert supplies data X of a process $\mathcal{E}$. Wants to know whether important relationships exist, and if so, what these relationships are.

Example areas: Oncology, Neurology, Brain Health.

2. Guidance of optimization algorithms

Example shown later: Dimension reduction of chemical process models.

3. (to be discovered – sorry, couldn't resist)

Summary: How to Find Significant Subgroups

Problem 1: Define target intervals I.

Solution:

Enumerate reasonable number of cases. Optionally, select cases by pattern analysis.

Problem 2: Find logic formula $R(x)$ for given target interval I.

Solution for the special case where each inequality has just one variable:

- Discretize the variables x_j.

- Formulate and solve an integer program (IP) whose solution allows separation of the discretized versions of the instances (x, t) with $t \in I$ from those with $t \notin I$. Tightly control the number of variables used in the IP solution.

- Translate the IP solution to a logic formula

$$R_1(x) \vee R_2(x) \vee \cdots R_k(x)$$

that separates the original instances (x, t) with $t \in I$ from those with $t \notin I$.

Each $R_i(x)$ is a conjunction of inequalities each of which has just one nonzero coefficient. Thus, the logic formula represents a union of rectangular polyhedra each of which potentially defines a subgroup.

Problem 3: Same as Problem 2, but the inequalities of $R_i(x)$ may have up to 4 nonzero coefficients.

Solution:

Expand X by adding variables y_j that are linear combinations of up to 4 x_j variables. Then use the solution method of Problem 2.

Problem 4: Construct logic formulas for which some $R_i(x)$ are significant with high probability and thus define significant subgroups.

Solution:

Evaluate *Alternate Random Processes* (ARPs) at each stage of the overall algorithm.

Application: Cervical Cancer

Data set supplied by the Frauenklinik, Charité, Berlin.

No prior information is given about goals of the analysis.

$n = 14$ variables
$N = 57$ cases of FIGO I-III cervical cancer

Table 1. Variables

Attribute	Uncertainty Interval
VEGF_PLASMA	[74.30 , 97.30]
VEGFD_SERUM	[381.00 , 441.00]
VEGFC_SERUM	[8455.00 , 9416.00]
ENDOGLIN	[4.06 , 4.63]
ENDOSTATIN	[123.00, 136.00]
ANGIOGENIN	[335.00 , 364.00]
FGFB_SERUM	[5.10 , 8.50]
VEGFR1_SERUM	[74.50 , 80.00]
VEGFR2_SERUM	[10995.00 , 11114.00]
M2PK_PLASMA	[20.80 , 21.80]
SICAM1_SERUM	[325.00 , 344.00]
SVCAM1_SERUM	[624.00 , 635.00]
IGFI_SERUM	[113.00 , 122.00]
IGFBP3_SERUM	[2552.00 , 2592.00]

Subgroup Discovery finds link between

-blood plasma/sera values measured from initial blood analysis and

- prediction whether treatment would ultimately be successful.

Rule:

If ENDOSTATIN < 123.0 or M2PK_PLASMA < 18.8,
then treatment most likely successful.

If ENDOSTATIN > 136.0 and M2PK_PLASMA > 21.8,
then treatment most likely not successful (cancer recurrence).

85% accuracy

Statistical significance: $p < 0.0002$

Application: Brain Injury of Children

Data supplied by Callier Center for Communication Disorders of U of Texas at Dallas.

Subgroup Discovery determines a lower bound connecting

(1) reduction of brain volume due to the injury
and
(2) the number of days till the patient has again a vocabulary of 10 words.

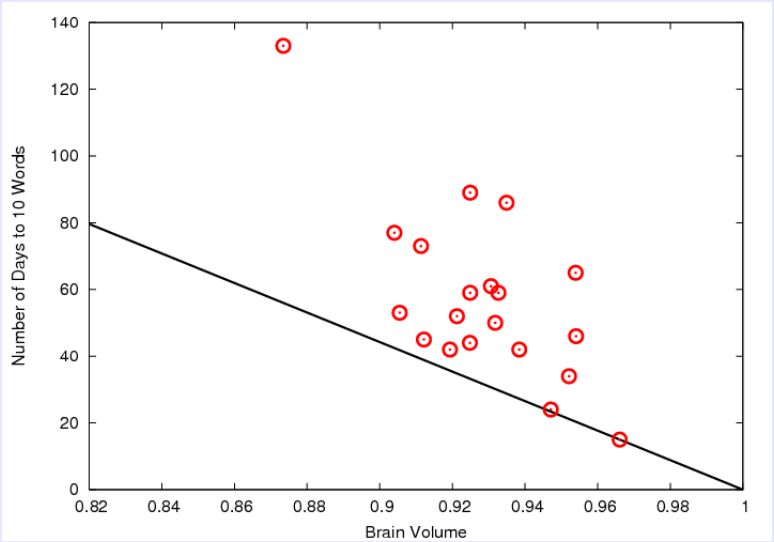


Fig. 1. Training Data: Brain Volume vs. Number of Days to 10 Words

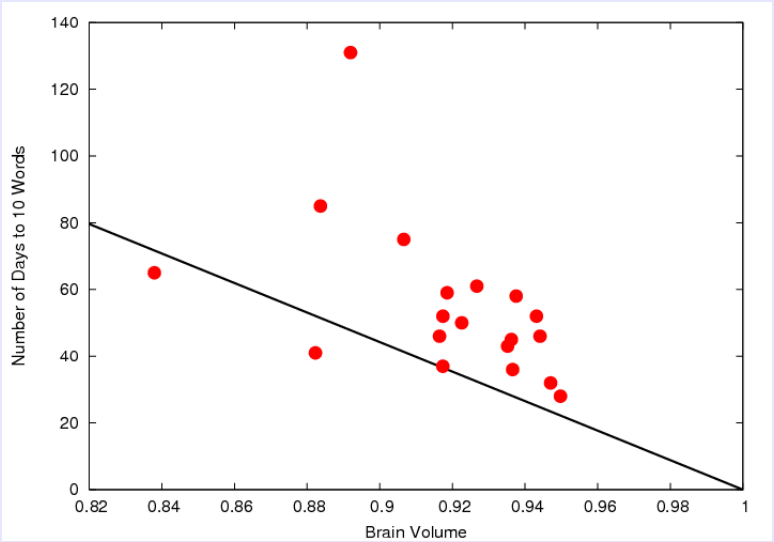


Fig. 2. Testing Data: Brain Volume vs. Number of Days to 10 Words

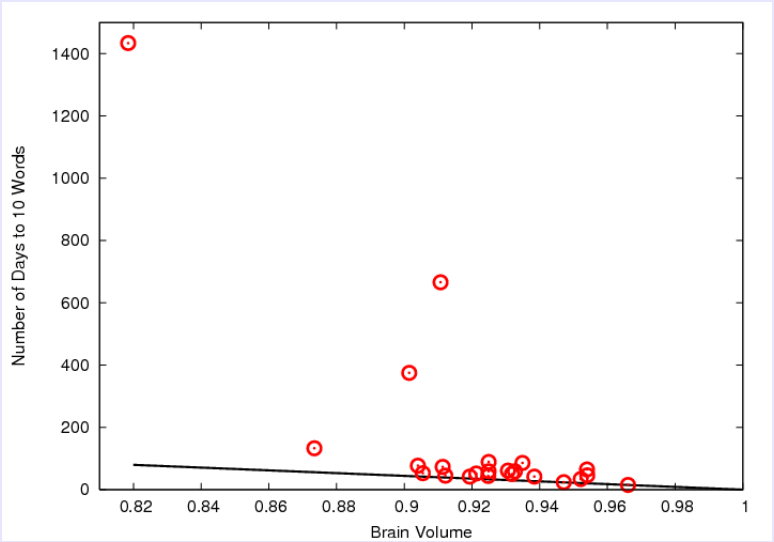


Fig. 3. All Training Data: Brain Volume vs. Number of Days to 10 Words

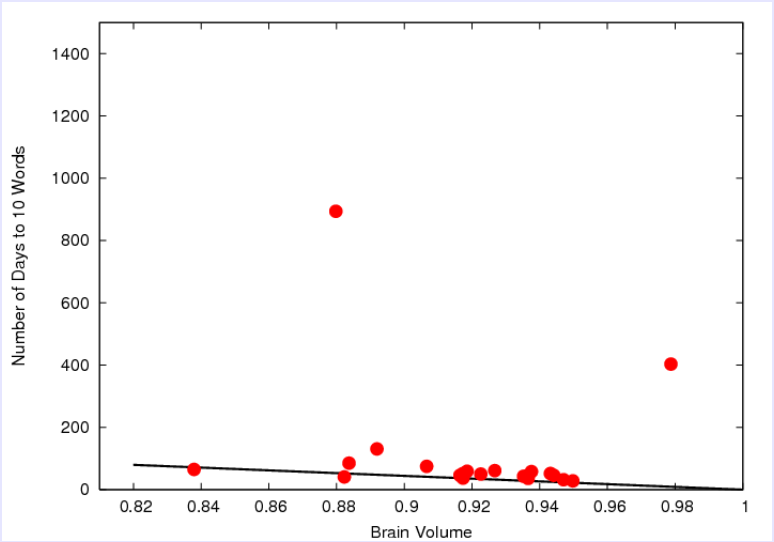


Fig. 4. All Testing Data: Brain Volume vs. Number of Days to 10 Words

Application: Classification of Children with Speech Delay

Problem: Characterize children with speech delay who do not respond to treatment.

Constitute about 10% of speech delay population.

Solution:

Find all important subgroups. For each subgroup, check if the characterization corresponds to a known classification. Any subgroup that does not correspond to a known classification and that has about 10% of the sample is a candidate for supplying the missing classification.

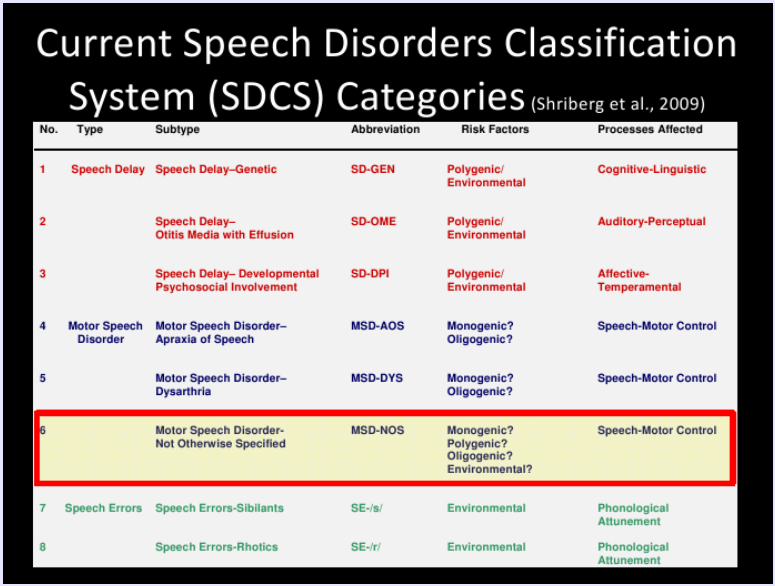


Fig. 5. Existing Classification

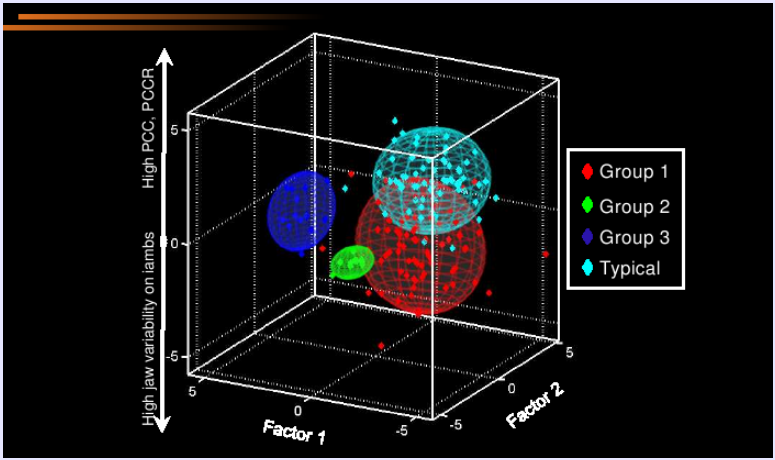


Fig. 6. Group 2 has size 9.7% and Likely Supplies Missing Classification

Dimension Reduction of Chemical Process Models

Work with G. Janiga, U of Magdeburg.

Process $\mathcal{E}$= Methane/air combustion.

Enthalpy of thermodynamic process
= total energy
$= U + pV$

where
U = internal energy
p = pressure at boundary
V = Volume

Vector x: 33 variables representing 29 gases, temperature, pressure, 2 velocity components
Function $F(x)$: enthalpy
Vector y: coordinates in plane where x vectors and $F(x)$ have been obtained.

Problem

Given: Simulation results = collection of $(x, F(x), y)$ vectors of combustion process $\mathcal{E}$.

Select a subvector z of the gases of x and a black box
such that

$\forall x = (z, z')$: the black box uses z to estimate z' and $F(x)$ with high accuracy.

Use of result: In similar settings where just z interaction is modeled, the black box estimates the z' values of x and $F(x)$.

Classical Solution Approach

Hoerl and Kennard (1970): "Ridge Regression" (2,339 citations)

Difficulty:

Must define nonlinear transformations for each x_j for reasonable representation of the behavior of x_j.

Assumptions

1. The given y vectors constitute a grid of a convex compact subset of R^m.

Assumption is trivially satisfied since the simulation creates data for a grid.

2. The function $F(x)$ is close to one-to-one for the given data.

Satisfied here since 3,655 vectors are given, and $F(x)$ has 3,412 distinct values.

Steps of Solution Method

1. Find highly significant subgroups for the x vectors, with $F(x)$ as target.

$\mathcal{I}$ = set of intervals I of the significant subgroups
P^I = polyhedron for case $I \in \mathcal{I}$

2. Compute significance measure q_j for each x_j.

Define *significance* q_j^I of x_j based on occurrence of x_j in the inequalities of P^I.

$q_j = \sum_{I \in \mathcal{I}} q_j^I$ = *overall significance* of x_j for $F(x)$.

Arguments for subsequent use of q_j:

- x_j with high q_j is important for computation of $F(x)$ values falling into some intervals $I \in \mathcal{I}$.

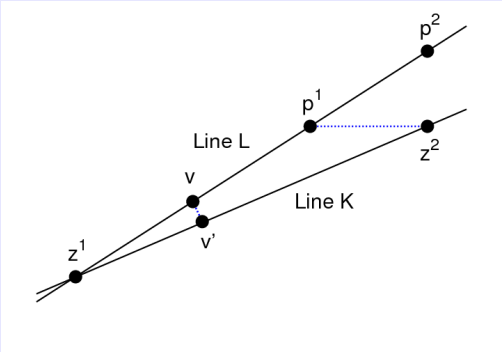
- Since $F(x)$ is almost one-to-one, x_j with high q_j is important for estimating x entries.

Hence: delete x_j only if q_j is small.

3. Define black box via a *Lazy Learner* and the given data $x = (z, z'), \forall x \in X$.

Input of black box: v
Output: v' for $x = (v, v')$

Method: Nearest Neighbor enhanced by interpolation.



Reduction Method

Recursive step:
1. Find overall significance values q_j. Let $q^* = \min_j q_j$.
2. For each x_j with q_j equal or close to q^*:
- delete x_j; use Lazy Learner to asses the estimation error for all variables deleted so far.
- Let x_{j^*} be index where error is minimum. Delete x_{j^*}.

Stop when error of any reduction exceeds a user-specified upper bound of the estimation error.

Methane /Air Combustion Application

33 variables: 29 gases, temperature, pressure, 2 velocity components.
Function: Enthalpy.

Algorithm reduces the 29 gases to 3 gases H_2, H_2O, and N_2.
The remaining variables and the enthalpy can be computed with rather good accuracy.

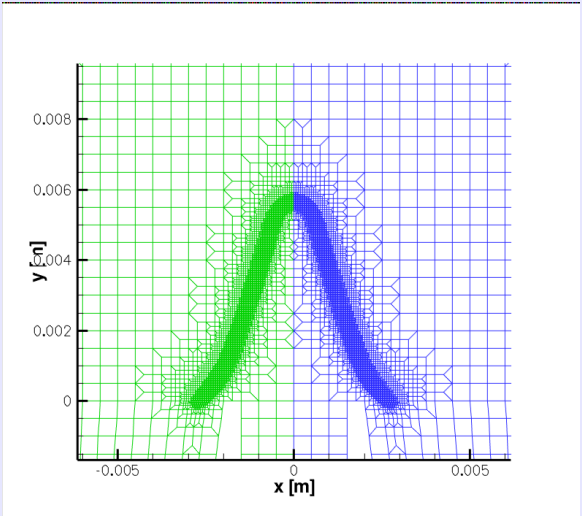


Fig. 7. Grid of Simulation Process

Table 2. Correlation of Actual and Estimated Values Using H_2, H_2O and N_2

Variable	Correlation	Variable	Correlation
u-velocity	0.9942	CH_2O	0.9998
v-velocity	0.9924	CH_3	0.9998
Pressure	0.9923	CH_3O	0.9996
Temperature	0.9989	CH_2OH	0.9998
H	0.9999	CH_4	0.9999
OH	0.9998	C_2H	0.9996
O	0.9998	HCCO	0.9996
HO_2	0.9996	C_2H_2	0.9998
H_2	1.0000	CH_2CO	0.9998
H_2O	1.0000	C_2H_3	0.9997
O_2	0.9999	C_2H_4	0.9997
CO	0.9999	C_2H_5	0.9997
CO_2	0.9999	C_2H_6	0.9997
CH	0.9997	C	0.9996
HCO	0.9998	C_2	0.9996
CH_2S	0.9998	N_2	1.0000
CH_2	0.9997	Enthalpy	0.9929

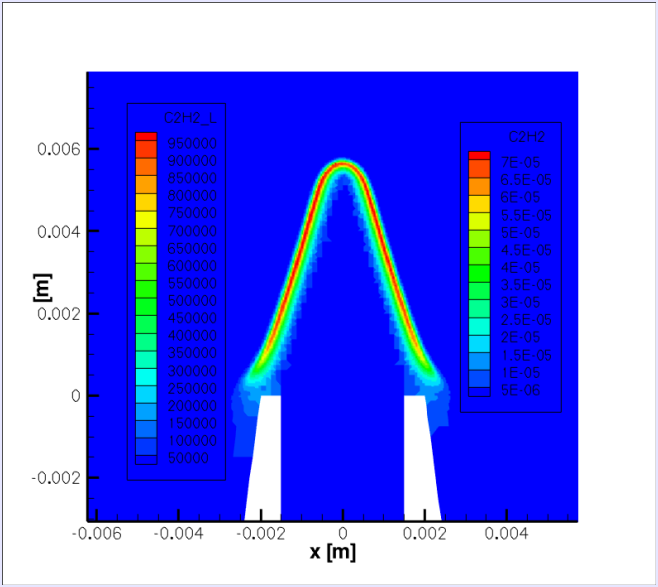


Fig. 8. 3-Variable Solution: Accuracy for C_2H_2

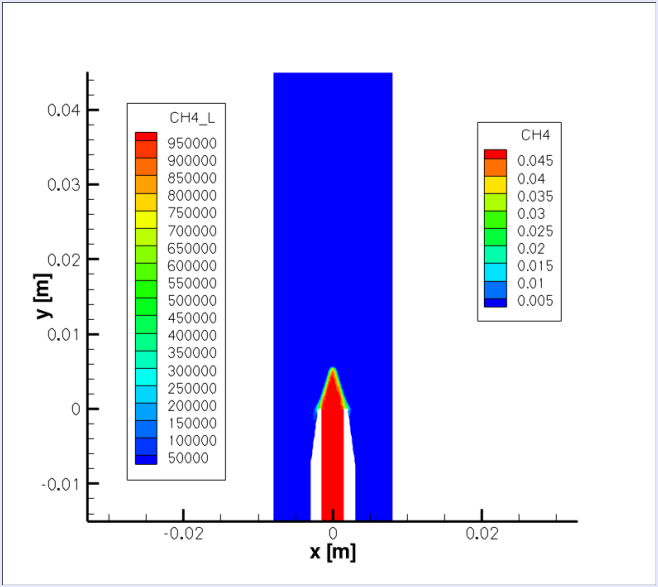


Fig. 9. 3-Variable Solution: Accuracy for CH_4

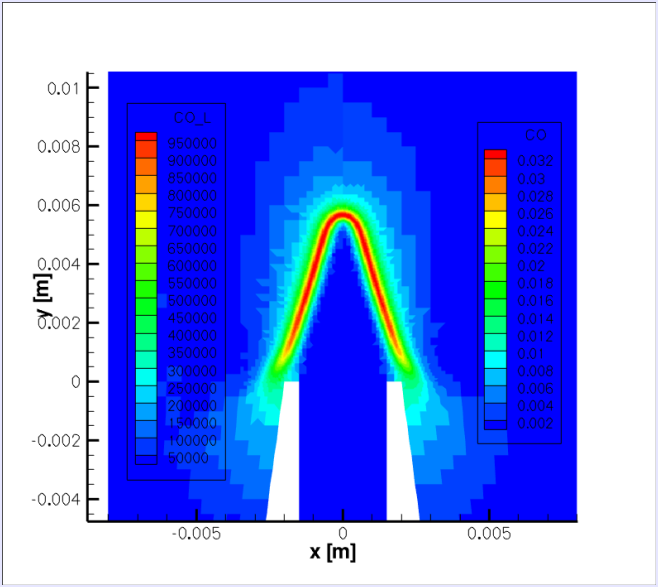


Fig. 10. 3-Variable Solution: Accuracy for CO

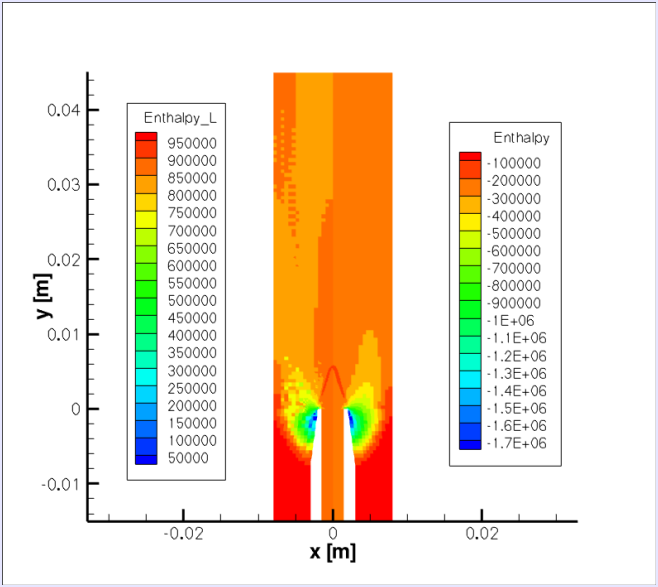


Fig. 11. 3-Variable Solution: Accuracy for Enthalpy

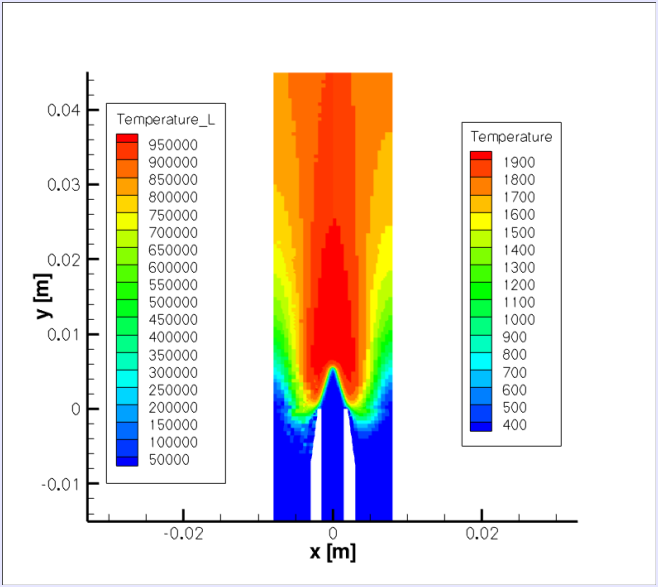


Fig. 12. 3-Variable Solution: Accuracy for Temperature

Similar, not quite as precise, results are obtained if the number of gases used for the estimation process is reduced to 2.

Comparison with Greedy and Optimal Solutions

Greedy solution excellent for 3 variables and poor for 2 variables.

Compared with the REDSUB solutions, the optimal solution has virtually the same accuracy for 3 variables, and is the same for 2 variables.