

# Améliorer la robustesse d'une sélection progressive de variables grâce à l'emploi de fichiers de données séparés

---

## Equation moyen infrarouge du méthane

**H. Soyeurt**, A. Vanlierde, C. Grelet, M.-L. Vanrobays, F. Colinet, N. Gengler, C. Bastin, F. Dehareng

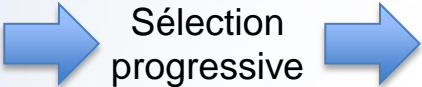


Gembloux Agro-Bio Tech  
Université de Liège



y

- variable 1
- variable 2
- variable 3
- variable 4
- variable 5
- variable 6
- variable 7
- variable 8
- variable 9
- .
- .
- .
- variable x



$$y = \text{variable ?} + \text{variable ?} + \dots + \text{variable ?}$$



y

- variable 1
- variable 2
- variable 3
- variable 4
- variable 5
- variable 6
- variable 7
- variable 8
- variable 9
- ⋮
- ⋮
- ⋮
- variable x

➔ Sélection progressive ➔

$$Y = \text{variable ?} + \text{variable ?} + \dots + \text{variable ?}$$



### INCONVENIENT

Peu robuste car équation liée aux spécificités du fichier de données de départ




y

- variable 1
- variable 2
- variable 3
- variable 4
- variable 5
- variable 6
- variable 7
- variable 8
- variable 9
- ⋮
- ⋮
- ⋮
- variable x

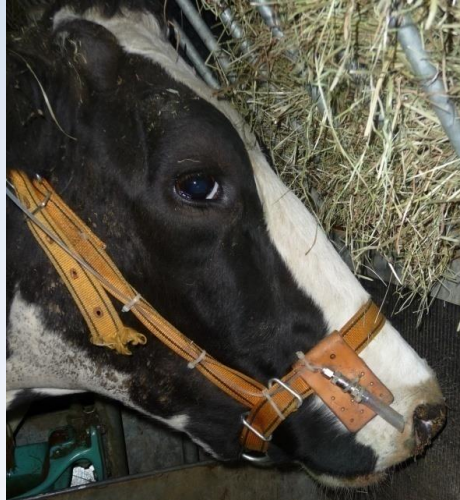
➔ Sélection progressive ➔

$$Y = \text{variable ?} + \text{variable ?} + \dots + \text{variable ?}$$

 **INCONVENIENT**  
Peu robuste car équation liée aux spécificités du fichier de données de départ

 **SOLUTION**  
Utiliser plusieurs fichiers de données et sélectionner des variables communes





**Méthode SF<sub>6</sub>**  
Coûteuse  
Contraignante



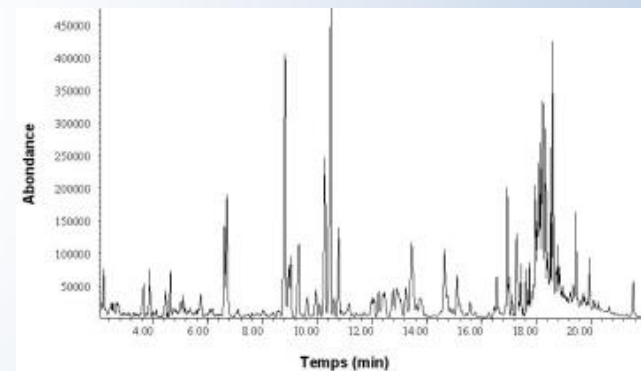
Gembloux Agro-Bio Tech  
Université de Liège



CH<sub>4</sub>

56 variables

Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique

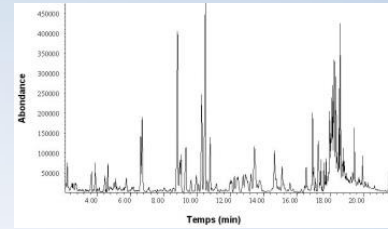


55 variables prédites par  
la spectrométrie MIR



Gembloux Agro-Bio Tech  
Université de Liège





56 variables dont 55 prédites par MIR

CH<sub>4</sub>

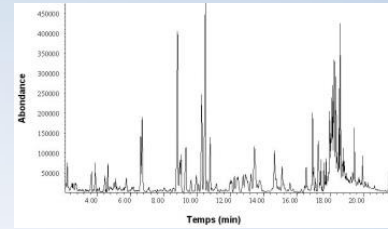
Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique

471 échantillons de lait + CH<sub>4</sub>



Gembloux Agro-Bio Tech  
Université de Liège





56 variables dont 55 prédits par MIR

CH<sub>4</sub>

Kg lait  
 %MG, %PROT  
 Acides gras (36)  
 Minéraux (5)  
 Compo protéique (4)  
 Propriétés fromagères (6)  
 Lactose, b-hydroxybutyrate,  
 acetone, acide citrique



CH<sub>4</sub> = .....



Données facilement  
 disponibles à faible  
 coût

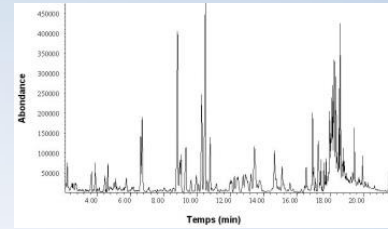
471 échantillons de lait + CH<sub>4</sub>



Gembloux Agro-Bio Tech  
 Université de Liège







56 variables dont 55 prédits par MIR

CH<sub>4</sub>

Kg lait  
 %MG, %PROT  
 Acides gras (36)  
 Minéraux (5)  
 Compo protéique (4)  
 Propriétés fromagères (6)  
 Lactose, b-hydroxybutyrate,  
 acetone, acide citrique



CH<sub>4</sub> = .....



Données facilement  
 disponibles à faible  
 coût

471 échantillons de lait + CH<sub>4</sub>



Gembloux Agro-Bio Tech  
 Université de Liège



CH<sub>4</sub>

Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique



CH<sub>4</sub> = .....

Fichier de données: séparation aléatoire

Set 1

Set 2

Set 3



Gembloux Agro-Bio Tech  
Université de Liège



CH<sub>4</sub>

Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique



CH<sub>4</sub> = .....

Fichier de données: séparation aléatoire

Set 1

Set 2

Set 3

**Calibration (2/3)**

**Validation (1/3)**

Set 1 + Set 2

Set 3

Set 2 + Set 3

Set 1

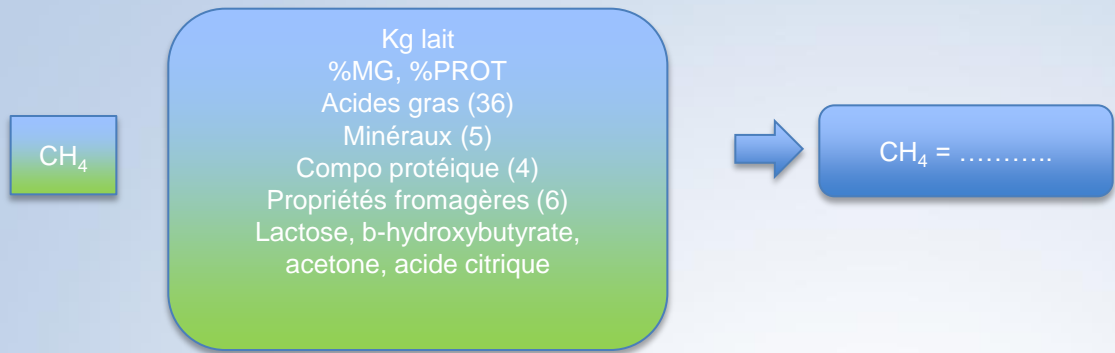
Set 1 + Set 3

Set 2



Gembloux Agro-Bio Tech  
Université de Liège

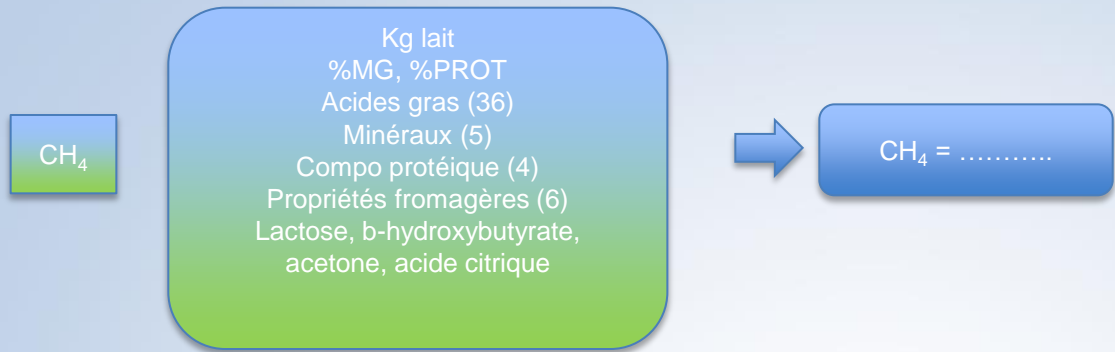




**SELECTION 1: PROC GLMSELECT(forward selection)**

➔ Schwarz Bayesian criterion (SBC) pour chaque variable

Variable	Calib 1	Variable	Calib 2	Variable	Calib 3
C18:2 c9,c12	2810.21	C18:2 c9,t11	2676.03	isoanteiso	2881.89
C18:2 c9,t11	2815.39	SCFA	2676.28	LCFA	2882.22
C18:3 c9,c12,c15	2815.95	isoanteiso	2677.26	omega6	2884.85
MCFA	2818.60	MCFA	2677.60	omega3	2886.49
tot c18:1 cis	2818.67	omega3	2677.77	SCFA	2889.23
SAT	2821.12	LCFA	2678.31	C18:2 c9,t11	2890.31
C18:2	2822.16	SAT	2681.47	ODD	2890.62
SCFA	2823.11	C18:2 c9,c12	2683.21	MCFA	2890.97
isoanteiso	2823.21	omega6	2683.38	C18:2 c9,c12	2892.86
LCFA	2823.84	ODD	2683.67	SAT	2896.09



**SELECTION 1: PROC GLMSELECT(forward selection)**

➔ Schwarz Bayesian criterion (SBC) pour chaque variable

Variable	Calib 1	Variable	Calib 2	Variable	Calib 3
C18:2 c9,c12	2810.21	C18:2 c9,t11	2676.03	isoanteiso	2881.89
C18:2 c9,t11	2815.39	SCFA	2676.28	LCFA	2882.22
C18:3 c9,c12,c15	2815.95	isoanteiso	2677.26	omega6	2884.85
MCFA	2818.60	MCFA	2677.60	omega3	2886.49
tot c18:1 cis	2818.67	omega3	2677.77	SCFA	2889.23
SAT	2821.12	LCFA	2678.31	C18:2 c9,t11	2890.31
C18:2	2822.16	SAT	2681.47	ODD	2890.62
SCFA	2823.11	C18:2 c9,c12	2683.21	MCFA	2890.97
isoanteiso	2823.21	omega6	2683.38	C18:2 c9,c12	2892.86
LCFA	2823.84	ODD	2683.67	SAT	2896.09

CH<sub>4</sub>

Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique



CH<sub>4</sub> = .....

SELECTION 1: PROC GLMSELECT(forward selection)

C18:2 c9,t11

SELECTION 2: PROC GLMSELECT(C18:2 c9,t11)

Schwarz Bayesian criterion (SBC) pour chaque ajout de variable

CH<sub>4</sub>

Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique



CH<sub>4</sub> = .....

SELECTION 1: PROC GLMSELECT(forward selection)

C18:2 c9,t11

SELECTION 2: PROC GLMSELECT(C18:2 c9,t11)

Schwarz Bayesian criterion (SBC) pour chaque ajout de variable

Variable	SBC	Variable	SBC	Variable	SBC
Caséine	2925.49	%PROT	2801.76	%PROT	3035.57
%PROT	2925.56	Ca	2807.93	%azote	3042.43
%azote	2932.87	%azote	2808.09	Caséine	3045.57
Ca	2939.83	Caséine	2817.35	Lactose	3055.41
Lactose	2955.99	Lactose	2817.55	Ca	3056.26
Mg	2967.80	Rdt_Sec	2830.27	Mg	3074.90
P	2990.37	Mg	2835.99	P	3089.34
Citrate	2994.07	Citrate	2850.01	Rdt_Sec	3095.81
Rdt_Sec	2994.47	P	2851.48	Citrate	3096.39
K	3023.50	Lait	2855.82	K	3110.88

CH<sub>4</sub>

Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique



CH<sub>4</sub> = .....

SELECTION 1: PROC GLMSELECT(forward selection)

C18:2 c9,t11

SELECTION 2: PROC GLMSELECT(C18:2 c9,t11)

Schwarz Bayesian criterion (SBC) pour chaque ajout de variable

Variable	SBC	Variable	SBC	Variable	SBC
Caséine	2925.49	%PROT	2801.76	%PROT	3035.57
%PROT	2925.56	Ca	2807.93	%azote	3042.43
%azote	2932.87	%azote	2808.09	Caséine	3045.57
Ca	2939.83	Caséine	2817.35	Lactose	3055.41
Lactose	2955.99	Lactose	2817.55	Ca	3056.26
Mg	2967.80	Rdt_Sec	2830.27	Mg	3074.90
P	2990.37	Mg	2835.99	P	3089.34
Citrate	2994.07	Citrate	2850.01	Rdt_Sec	3095.81
Rdt_Sec	2994.47	P	2851.48	Citrate	3096.39
K	3023.50	Lait	2855.82	K	3110.88



CH<sub>4</sub>

Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique



CH<sub>4</sub> = .....

SELECTION 1: PROC GLMSELECT(forward selection)

C18:2 c9,t11

SELECTION 2: PROC GLMSELECT(C18:2 c9,t11)

%PROT

Schwarz Bayesian criterion (SBC) pour chaque ajout de variable



SELECTION XX: PROC GLMSELECT (variables sélectionnées)

Ici 13 variables

## CH<sub>4</sub> (g/jour)

C18:2 c9,t11

%PROT

Milk yield

A30

Omega 3

Total trans FA

C18:2 c9,c12

Iso + Anteiso FA

C15:0 anteiso

C18:3 c9,c12,c15

%azote

Acetone

R

Calib 1

Calib 2

Calib 3

**Root MSE**

**72.07**

**67.99**

**71.21**

Mean

417.90

423.74

421.33

**R<sup>2</sup>**

**0.97**

**0.98**

**0.98**

SBC

2773.59

2592.51

2834.62

**ASE (Train)**

**4980.68**

**4422.80**

**4868.56**

**ASE (Test)**

**4770.19**

**5701.51**

**5001.75**

# CH<sub>4</sub> (g/jour)

C18:2 c9,t11

%PROT

Milk yield

A30

Omega 3

Total trans FA

C18:2 c9,c12

Iso + Anteiso FA

C15:0 anteiso

C18:3 c9,c12,c15

%azote

Acetone

R

Calib 1

Calib 2

Calib 3

**Root MSE**

**72.07**

**67.99**

**71.21**

Mean

417.90

423.74

421.33

**R<sup>2</sup>**

**0.97**

**0.98**

**0.98**

SBC

2773.59

2592.51

2834.62

**ASE (Train)**

**4980.68**

**4422.80**

**4868.56**

**ASE (Test)**

**4770.19**

**5701.51**

**5001.75**

**Root MSE**

**72.10**

**67.95**

**71.12**

Mean

417.90

423.74

421.33

**R<sup>2</sup>**

**0.69**

**0.73**

**0.73**

SBC

2778.59

2596.82

2838.50

**ASE (Train)**

**4968.87**

**4402.40**

**4840.07**

**ASE (Test)**

**4703.63**

**5670.49**

**5006.34**



Calibration 1			Calibration 2			Calibration 3		
Lait	<b>RMSE</b>	<b>71.48</b>	%PROT	<b>RMSE</b>	<b>66.52</b>	Lait	<b>RMSE</b>	<b>70.30</b>
omega3	Mean	417.90	Lait	Mean	423.74	A30	Mean	421.33
Total trans	<b>R<sup>2</sup></b>	<b>0.97</b>	Total trans	<b>R<sup>2</sup></b>	<b>0.98</b>	Total trans	<b>R<sup>2</sup></b>	<b>0.98</b>
R	SBC	2763.70	C16:0	SBC	2565.36	isoanteiso	SBC	2826.18
C14:1	<b>ASE</b>	<b>4916.18</b>	C17:0	<b>ASE</b>	<b>4277.25</b>	C18:3	<b>ASE</b>	<b>4743.80</b>
	(Train)			(Train)		c9,c12c15	(Train)	
SCFA	<b>ASE</b>	<b>5722.55</b>	C18:0	<b>ASE</b>	<b>6096.35</b>	%azote	<b>ASE</b>	<b>5248.22</b>
	(Test)			(Test)			(Test)	
C15:0			SCFA			Acetone		
Total C15			LCFA			C14:0		
Ca			BHB			C18:2		
K			Citrate			MCFA		
Urée						omega6		
Caséines						BHB		
						Citrate		

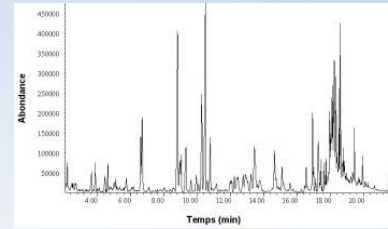


Calibration 1			Calibration 2			Calibration 3		
Lait	<b>RMSE</b>	<b>71.48</b>	%PROT	<b>RMSE</b>	<b>66.52</b>	Lait	<b>RMSE</b>	<b>70.30</b>
omega3	Mean	417.90	Lait	Mean	423.74	A30	Mean	421.33
Total trans	<b>R<sup>2</sup></b>	<b>0.97</b>	Total trans	<b>R<sup>2</sup></b>	<b>0.98</b>	Total trans	<b>R<sup>2</sup></b>	<b>0.98</b>
R	SBC	2763.70	C16:0	SBC	2565.36	isoanteiso	SBC	2826.18
C14:1	<b>ASE</b>	<b>4916.18</b>	C17:0	<b>ASE</b>	<b>4277.25</b>	C18:3	<b>ASE</b>	<b>4743.80</b>
	(Train)			(Train)		c9,c12c15	(Train)	
SCFA	<b>ASE</b>	<b>5722.55</b>	C18:0	<b>ASE</b>	<b>6096.35</b>	%azote	<b>ASE</b>	<b>5248.22</b>
	(Test)			(Test)			(Test)	
C15:0			SCFA			Acetone		
Total C15			LCFA			C14:0		
Ca			BHB			C18:2		
K			Citrate			MCFA		
Urée						omega6		
Caséines								

<b>Root MSE</b>	<b>72.0</b>	<b>67.99</b>	<b>71.21</b>
<b>Mean</b>	417.90	423.74	421.33
<b>R<sup>2</sup></b>	<b>0.97</b>	<b>0.98</b>	<b>0.98</b>
<b>SBC</b>	2773.59	2592.51	2834.62
<b>ASE (Train)</b>	<b>4980.68</b>	<b>4422.80</b>	<b>4868.56</b>
<b>ASE (Test)</b>	<b>4770.19</b>	<b>5701.51</b>	<b>5001.75</b>



# Conclusions



56 variables dont 55 prédits par MIR

CH<sub>4</sub>

Kg lait  
%MG, %PROT  
Acides gras (36)  
Minéraux (5)  
Compo protéique (4)  
Propriétés fromagères (6)  
Lactose, b-hydroxybutyrate,  
acetone, acide citrique



CH<sub>4</sub> = C18:2 c9,t11 + %PROT  
+ Lait + A30 + Omega 3 + Total  
trans FA + C18:2 c9,c12 +  
Iso\_Anteiso FA + C15:0  
anteiso + C18:3 c9,c12,c15 +  
%azote + Acetone + R



Données facilement  
disponibles à faible  
coût

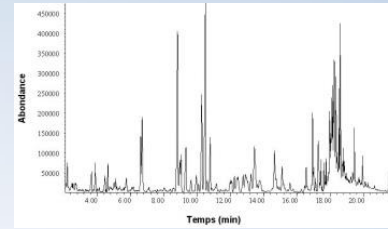
471 échantillons de lait + CH<sub>4</sub>



Gembloux Agro-Bio Tech  
Université de Liège



# Conclusions



56 variables dont 55 prédits par MIR

CH<sub>4</sub>

Kg lait  
 %MG, %PROT  
 Acides gras (36)  
 Minéraux (5)  
 Compo protéique (4)  
 Propriétés fromagères (6)  
 Lactose, b-hydroxybutyrate,  
 acetone, acide citrique



CH<sub>4</sub> = C18:2 c9,t11 + %PROT  
 + Lait + A30 + Omega 3 + Total  
 trans FA + C18:2 c9,c12 +  
 Iso\_Anteiso FA + C15:0  
 anteiso + C18:3 c9,c12,c15 +  
 %azote + Acetone + R



Données facilement  
 disponibles à faible  
 coût

10 x	N calib	N valid	ASE calib	ASE valid	R <sup>2</sup> calib	R <sup>2</sup> valid
Mean	325	146	4728.24	5295.07	<b>0.72</b>	<b>0.68</b>
SD	9	9	191.94	517.94	<b>0.01</b>	<b>0.03</b>
Min	306	135	4517.35	4395.57	<b>0.71</b>	<b>0.63</b>
Max	336	165	5043.48	6105.56	<b>0.74</b>	<b>0.71</b>

# Améliorer la robustesse d'une sélection progressive de variables grâce à l'emploi de fichiers de données séparés

---

## Equation moyen infrarouge du méthane

**H. Soyeurt**, A. Vanlierde, C. Grelet, M.-L. Vanrobays, F. Colinet, N. Gengler, C. Bastin, F. Dehareng



Gembloux Agro-Bio Tech  
Université de Liège

