DATAMINING & GENESIS

Pourquoi le Datamining ?

> Les administrations fiscales disposent d'une grande quantité de données (personnelles, fiscales, etc.)

> Disponibles sur support papier ou dans les bases de données (datawarehouse, datamarts)

> Base de données utilisées principalement pour la consultation

> Analyse de risque inexistante ou minimale (ou mal organisée) :

- l'administration fiscale appliquait l'analyse de risque d'une manière spontanée au niveau local ➔ à harmoniser par des outils au niveau central pour atteindre un traitement uniforme
- des listes de sélection de toute nature et composition étaient déjà diffusées dans tous les services, sans cependant avoir une vision générale sur une gestion de risque ➔ à harmoniser dans un programme de contrôle (avec utilisation de STIR CO)

Outils utilisés pour le Datamining dans l'administration fiscale belge

> Clementine (SPSS)/SAS Miner + techniques datamining :

- Migration complète en 2011 vers SAS Miner + SAS Enterprise Guide

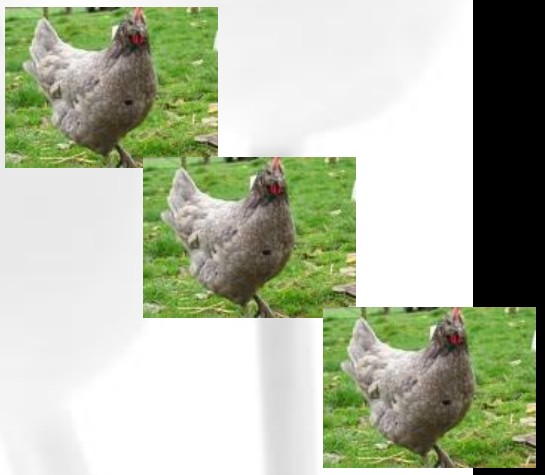
> Un développement évolutif

- Modèles "Fraude classique à la TVA"
- Modèles pour sélection des remboursements TVA (L678)
- Modèles de détection pour les fraudes carrousel

Qu'est-ce que le datamining ?

Utiliser les données du passé …

… Pour prédire les événements du futur !



Avantages du datamining

> Une très grande quantité de données (humainement impossibles à interpréter) : trier un très grand nombre de données relatives à +/- 650.000 entreprises n'est pas réalisable sans l'aide d'un outil datamining

> Un modèle dynamique : le modèle peut être ajusté sur base annuelle/trimestrielle/mensuelle

> Multiples possibilités d'application : fraude classique, opérateurs défaillants, remboursements TVA, risques liés aux nouveaux assujettis, etc..

Méthodologie

- Compréhension du “Business”
- Compréhension des données
- Préparation des données
- Modélisation
- Evaluation
- Déploiement
- Autres méthodes (SAS) = SEMMA (Sample, Explore, Modify, Model, Assess)

Modèle DM TVA

> Le modèle TVA est simplement basé sur une confrontation des :

> Suppléments TVA historiques >< Données explicatives

> OU = “passé de la fraude” vs “l’entièreté des données concernant les cas de fraude détectés dans le passé”

Gestion du projet Datamining : Méthodologie CRISP

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
DM Goals
DM Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

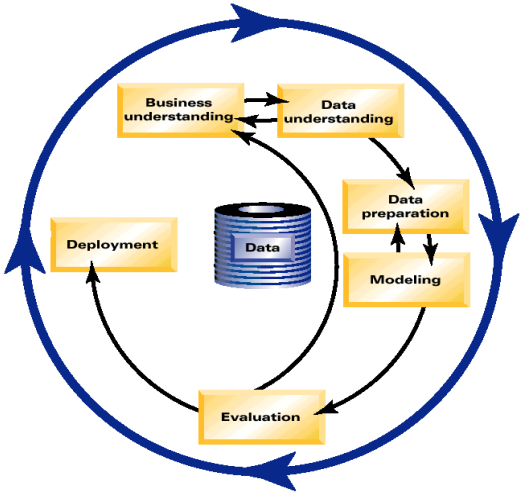
Deployment

Plan Deployment
Deployment Plan

Plan Monitoring And Maintenance
Monitoring and Maintenance Plan

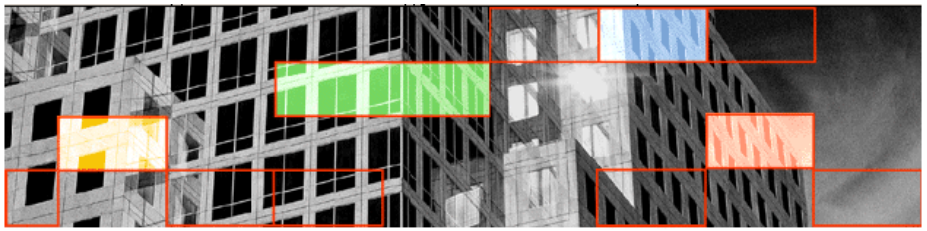
Produce Final Reportl
Final Report
Final Presentation

Review Project
Experience Documentation



Méthodologie

- **COMPREHENSION DU "BUSINESS"**
- **COMPREHENSION DES DONNEES**
  - Collecte des données
  - Etude qualitative des données
- **PREPARATION DES DONNEES**
  - Mise en forme des données
  - Nettoyage des données
  - Définition du Datamart : MS SQL Server 2000
- **MODELISATION**
- **EVALUATION**
- **DEPLOIEMENT**



Objectifs du "Business" et Critères de Succès

> Maximiser le recouvrement TVA + impôts directs ➔ absence de déclaration de la totalité du chiffre d'affaire – déduction exagérée (fraude générale à la TVA) vs. Fraude MTIC (grande importance monétaire + occurrence unique)

> Maximiser la productivité de l'administration fiscale

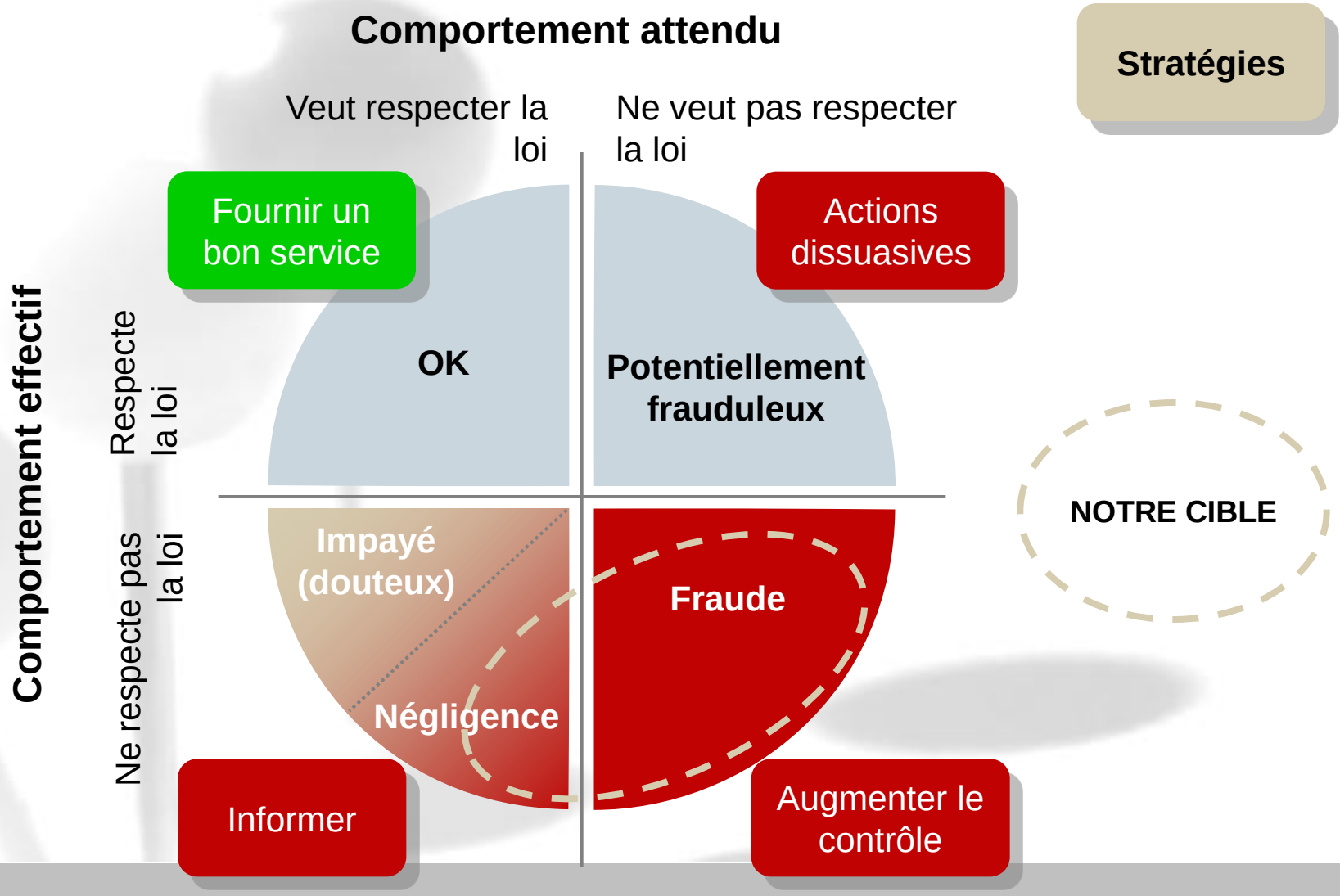
> Appliquer des contrôles et des standards de sélection de manière uniforme pour tous les assujettis : principe de traitement égal

> Soutenir des actions d'investigations nationales pour des groupes cibles spécifiques

> Permettre des choix stratégiques (ex. ressources humaines : tissus fiscal)

> Possibilité de découvrir de nouvelles techniques (jusqu'alors inconnues) pour détecter les assujettis qui commettent des fraudes TVA

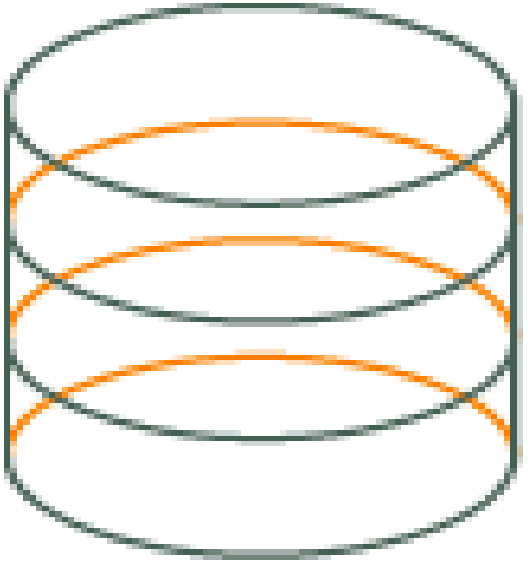
Visualisation de la cible "fraude classique"



Analyse du "Business"

> La compréhension et l'analyse du "Business" permettent aux techniciens de reconnaître :
  - les données utilisées par les contrôleurs ou les agents qui explorent les données dans le but d'obtenir des sélections manuelles
  - les indicateurs utilisés par les contrôleurs ou les agents qui font les sélections manuelles (contexte logique des données)
  - l'importance relative (pertinence) des indicateurs (approprié pour quel type de fraude, quel secteur d'activité et dans quelle proportion)
> Dans l'itération actuelle du Datamining (DM8), on a utilisé 159 indicateurs pour 11 modèles pour la fraude classique TVA et pour 1 modèle pour les remboursements à la TVA
> Dans le cadre VAT-Package : 26 nouveaux indicateurs développés
> Historique des données ➔ limité de 1995 vers 2005 jusqu'à présent

Données disponibles >< Données utilisables



> Déclarations périodiques à la TVA (X)

> Déclarations fiscales ISR

> Listing IntraCommunautaires (X)

> Listing Fournisseurs (X)

> VIES (X)

> Dettes fiscales – Bilan fiscal

> Compte courant TVA (X)

> Résultats de contrôles antérieurs de l'administration fiscale (X)

> Données sur les propriétés immobilières des contribuables

> Véhicules immatriculés par les contribuables

> ...

Compréhension et préparation des données

> Choix des données appropriées : compréhension du contenu des bases de données

> Données récentes : limitées pour la production aux données de la dernière année disponible (càd 2004 pour DM2, 2006 pour DM4, 2008 pour DM6) et aux données historiques de la période 2005 jusqu'au présent pour la modélisation

> Données utilisables (qualité) : nettoyage ou élimination des bases de données inappropriées : permutation des données vers des valeurs logarithmiques pour écarter les valeurs extrêmes (GE)

> Données agregées : toutes les confrontations de données se font sur base d'un an

Définition des segments homogènes

Splitting:

➔Très grandes entreprises

➔CA très bas

- Industrie
- Construction
- Automobiles
- Grossistes
- Détaillants
- Hotels/restaurants/cafés
- Services : ICT/Consultance >< Autres services

Mixité

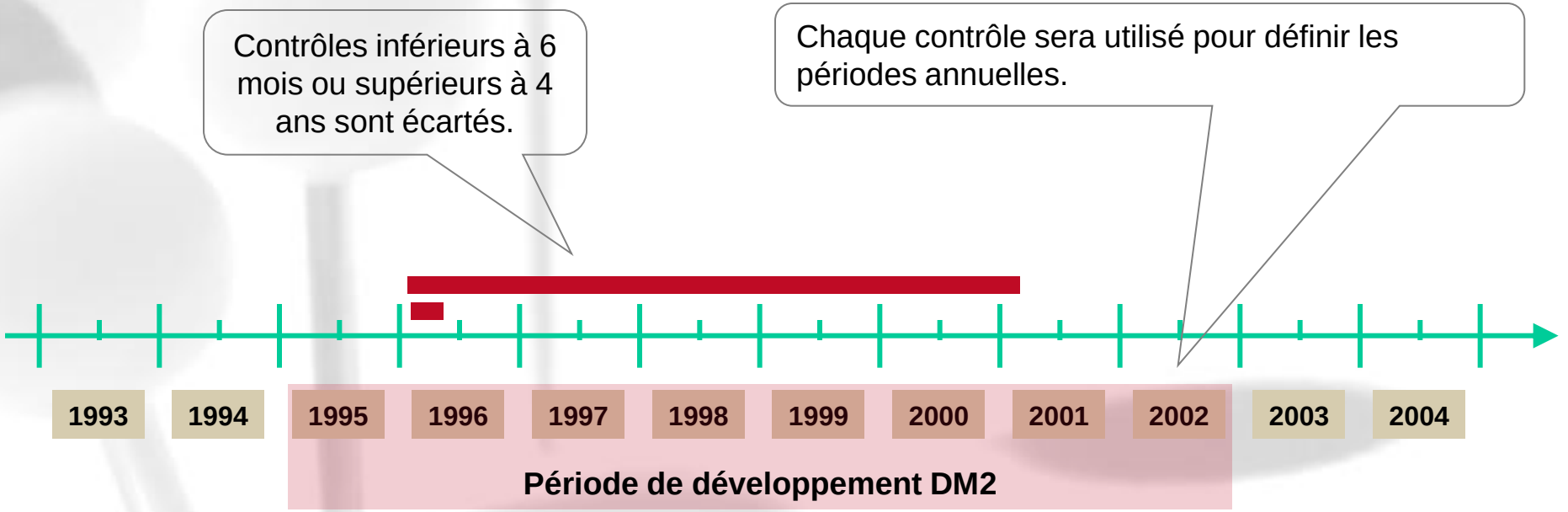
Active filing : AM AT LT

Entreprises art. 4 Code TVA

Définition de la période de contrôle

Fraude «Classique»

> Référence de temps (les champs “date” dans la base de données doivent être valides)



Modélisation en 2 étapes

1. Etape Développement :

- Répartition 50/50 % des cas historiques sur : Entraînement >< Test
- Repose sur la discrimination entre les résultats “positifs” et “négatifs” en utilisant différents seuils (= cut-off) sur les données historiques de “périodes contrôlées de 1 an”. De là, on développe un modèle pour chaque segment en utilisant un des algorithmes suivants:

➢ *Le réseau neuronal* : en mesurant les caractéristiques ou indicateurs en fonction de leur pertinence selon les données historiques, on aboutit à un “scoring” des assujettis de toute la population selon les données récentes

➢ *L’arbre de décision* : une discrimination continue mène à l’isolement, d’un point de vue statistique, des membres les plus intéressants d’une population.

Nombreuses variétés disponibles : C5.0; CHAID; CART, etc…

➢ *Autres* : régression linéaire ou logistique …

2. Etape Production :

- On applique le modèle sur la population entière des contribuables

Stratégie de modélisation

> **Booléen** (bipolaire : vrai/faux, cela nous donne des cas positifs ou négatifs)

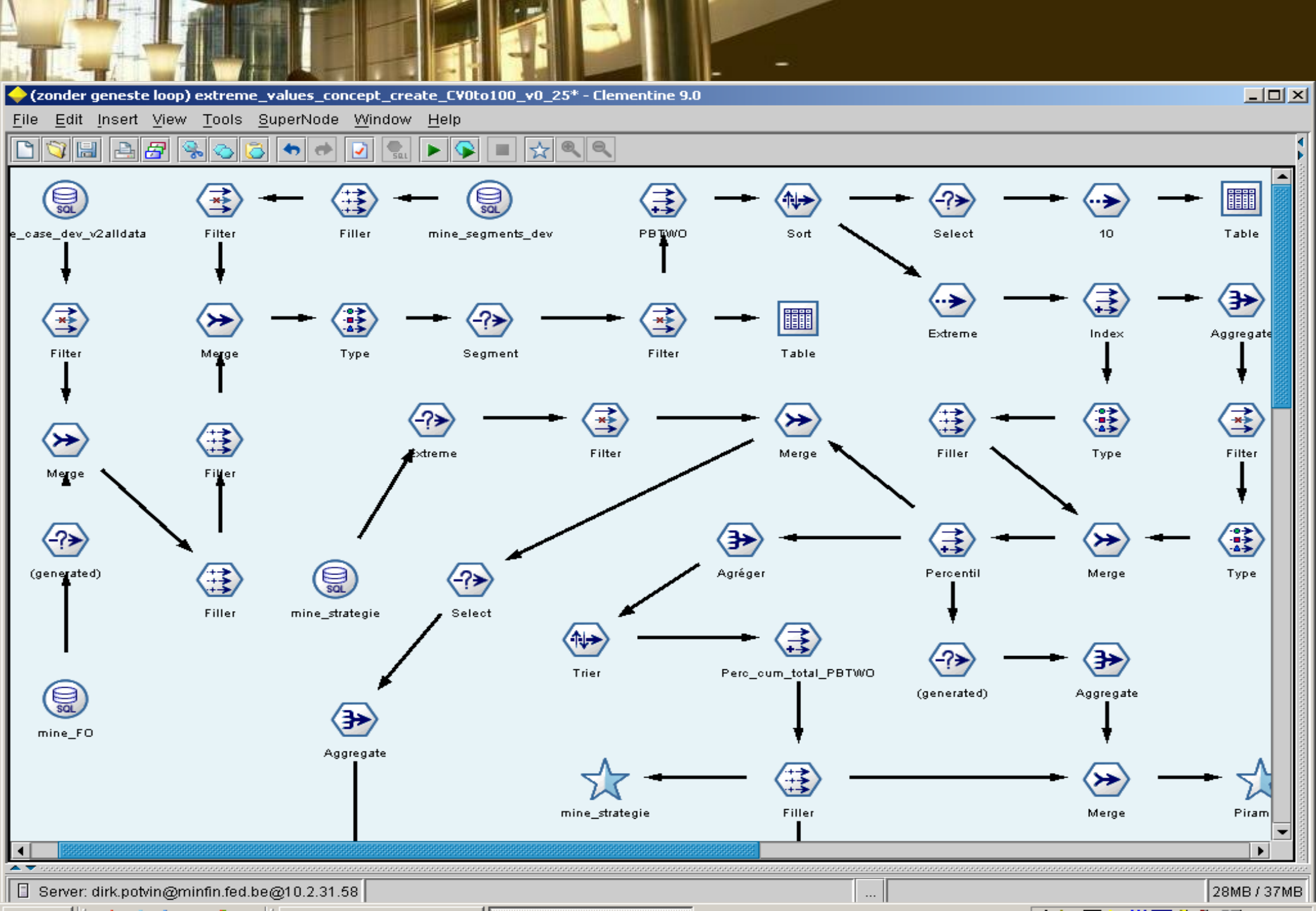
> **Cible** (cut off) :

- Horeca : 11.699 €
- Mixité TVA : 9.370 €
- Véhicules : 6.697 €
- Industrie : 5.342 €
- Grossistes : 4.909 €
- Services : 4.408 €
- Construction : 3.814 €

DM pour la fraude générale à la TVA :

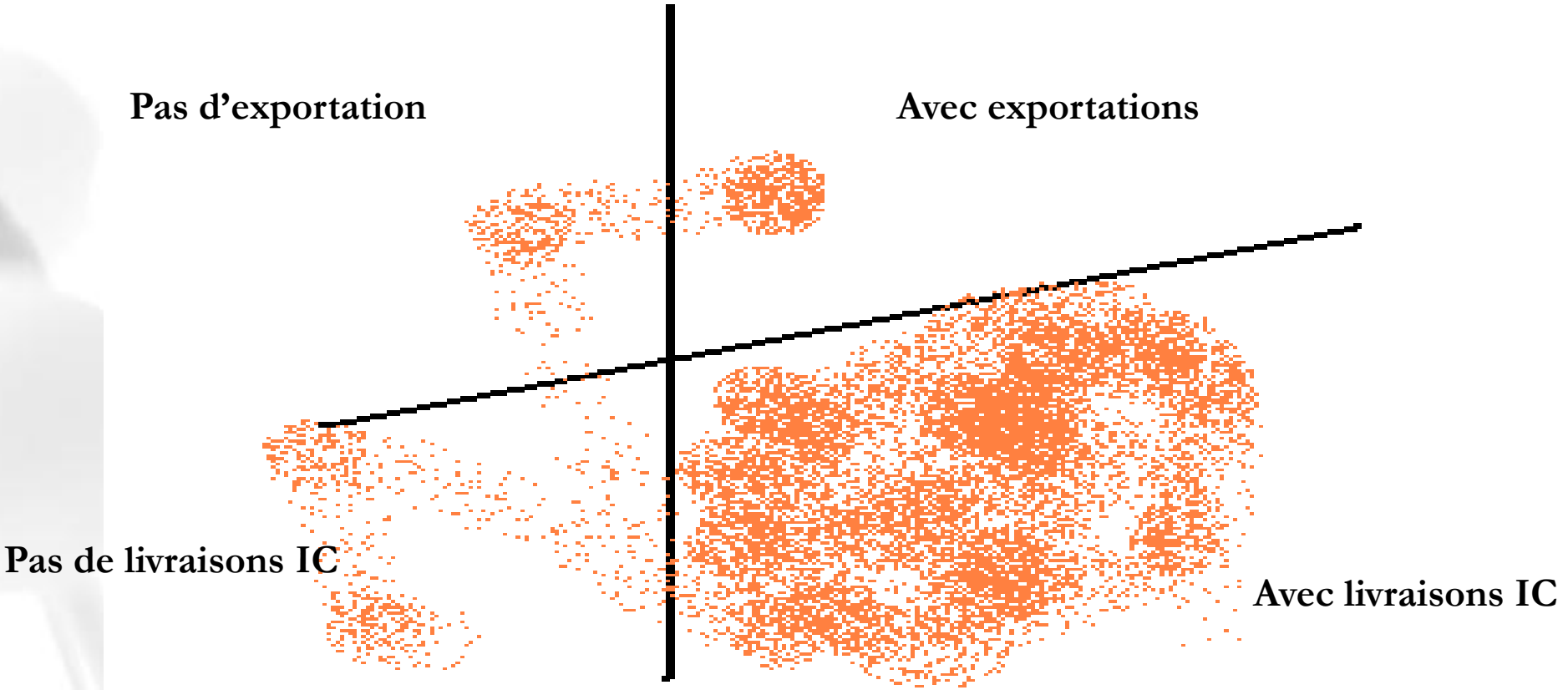
> Utilisation des indicateurs Expert (competitors) pour le segment («Liste Expert») + variables d'input complémentaires sélectionnés par le logiciel

> Valeurs sans transformation logarithmique : petites et grandes entreprises


(zonder geneste loop) extreme_values_concept_create_CV0to100_v0_25* - Clementine 9.0
File Edit Insert View Tools SuperNode Window Help
e_case_dev_v2alldata
Filter
Filler
mine_segments_dev
PBTWO
Sort
Select
10
Table
Filter
Merge
Type
Segment
Filter
Table
Extreme
Index
Aggregate
Merge
Filler
Extreme
Filter
Merge
Filler
Type
Filter
(generated)
Filler
mine_strategie
Select
Agréger
Percentil
Merge
Type
Trier
Perc_cum_total_PBTWO
(generated)
Aggregate
mine_FO
Aggregate
mine_strategie
Filler
Merge
Piram
Server: dirk.potvin@minfin.fed.be@10.2.31.58
28MB / 37MB

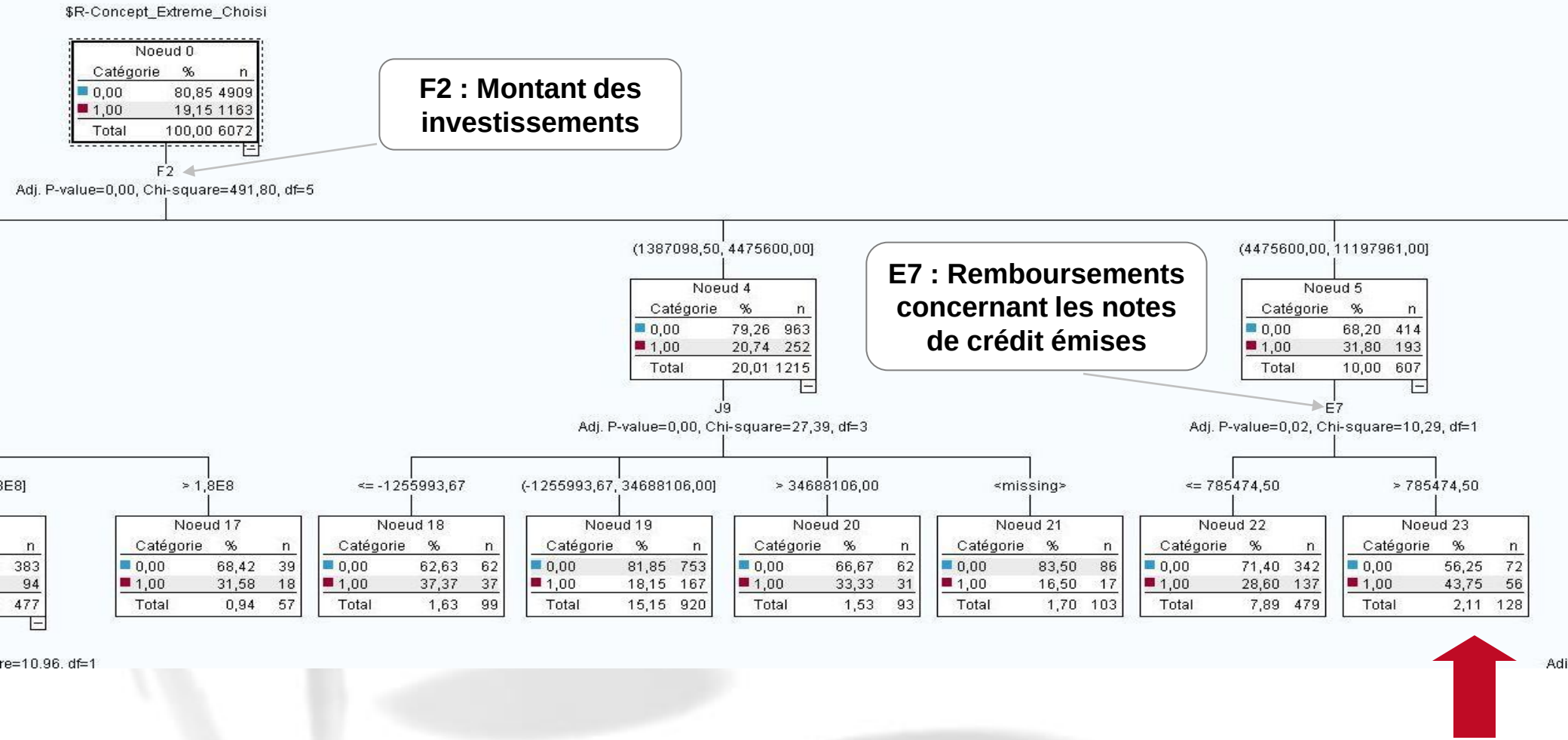
```
•IF  NOT MISSING(I1 ) AND
  I1  <          296018.625 THEN DO;
  IF  NOT MISSING(D4 ) AND
    D4  <   2.2306329682129E-6 THEN DO;
    IF  NOT MISSING(S1 ) AND
      S1  <     489.089140914262 THEN DO;
      IF  NOT MISSING(V5 ) AND
        V5  <    -2.67699329545627 THEN DO;
        IF  NOT MISSING(B7 ) AND
            0.00234493836947 <= B7  AND
          B7  <     0.99835332615807 THEN DO;
          _NODE_  =                 57;
          _LEAF_  =                 2;
          P_Concept_Extreme_Choisi1  =     0.23905723905723;
          P_Concept_Extreme_Choisi0  =     0.76094276094276;
          V_Concept_Extreme_Choisi1  =     0.22948073701842;
          V_Concept_Extreme_Choisi0  =     0.77051926298157;
          I_Concept_Extreme_Choisi  = '0' ;
          U_Concept_Extreme_Choisi  = '0' ;
          END;
```

Exemple de discrimination



ARBRE DE DECISION : exemple

Arbre de décision : fonctionnement



Mathématiques derrière le modèle : test CHI²

Neyman-Pearson ➔ test *le plus puissant* : rapport de vraisemblance *chi²*

test CHI² : $X^2 = \sum (N - e)^2/e$ – Population = 60

N = valeur détectée : $n^1=13$; $n^2=9$; $n^3=8$; $n^4=11$; $n^5=5$; $n^6=14$

E = valeur attendue = 10

$X^2 = (3^2 + 1^2 + 2^2 + 1^2 + 5^2 + 4^2)/10 = (9 + 1 + 4 + 1 + 25 + 16)/10 = 5,6$

- En prenant en compte une déviance maximale de la valeur (=5), on peut calculer la probabilité (p) que la déviance constatée se manifestera (en fonction du nombre de cas)
- Dans nos modèles, on accepte uniquement : p > 95 % ➔ cela implique qu'il est sûr à 95% que cette déviance calculée sera confirmée dans la réalité (! ≠ 95% de "cas positifs" !!!)

Peaufiner le modèle : coût des erreurs de classification

> Le **coût des erreurs de classification (= coût si "faux négatif")** n'est pas pertinent durant la phase de développement du modèle

> Mais lorsque le modèle est appliqué sur la partition de test dans le but de prédire la cible, le **coût des erreurs de classification** est utilisé (au plus le coût est élevé => moins grand nombre de faux négatifs)

- Horeca : 6,2
- Mixité TVA : 4,6
- *Détaillants* : 4,6
- Véhicules : 4,5
- Industrie : 4,5
- Services : 4,3
- *Construction* : 4,3
- Grossistes : 4
- Activités avec un faible chiffre d'affaires : 4

Pourquoi de si petits chiffres ? Augmentation du **coût des erreurs de classification** ➔ amélioration de la performance, mais au détriment de la stabilité du modèle (résultats variables)

Remarque : 100 pour la fraude MTIC ➔ pas de faux négatifs !

Evaluation du modèle

> LEVIER :

- Le gain que la sélection d'un point de vue mathématique permet d'obtenir en utilisant le modèle au lieu d'une sélection aléatoire (ad random)

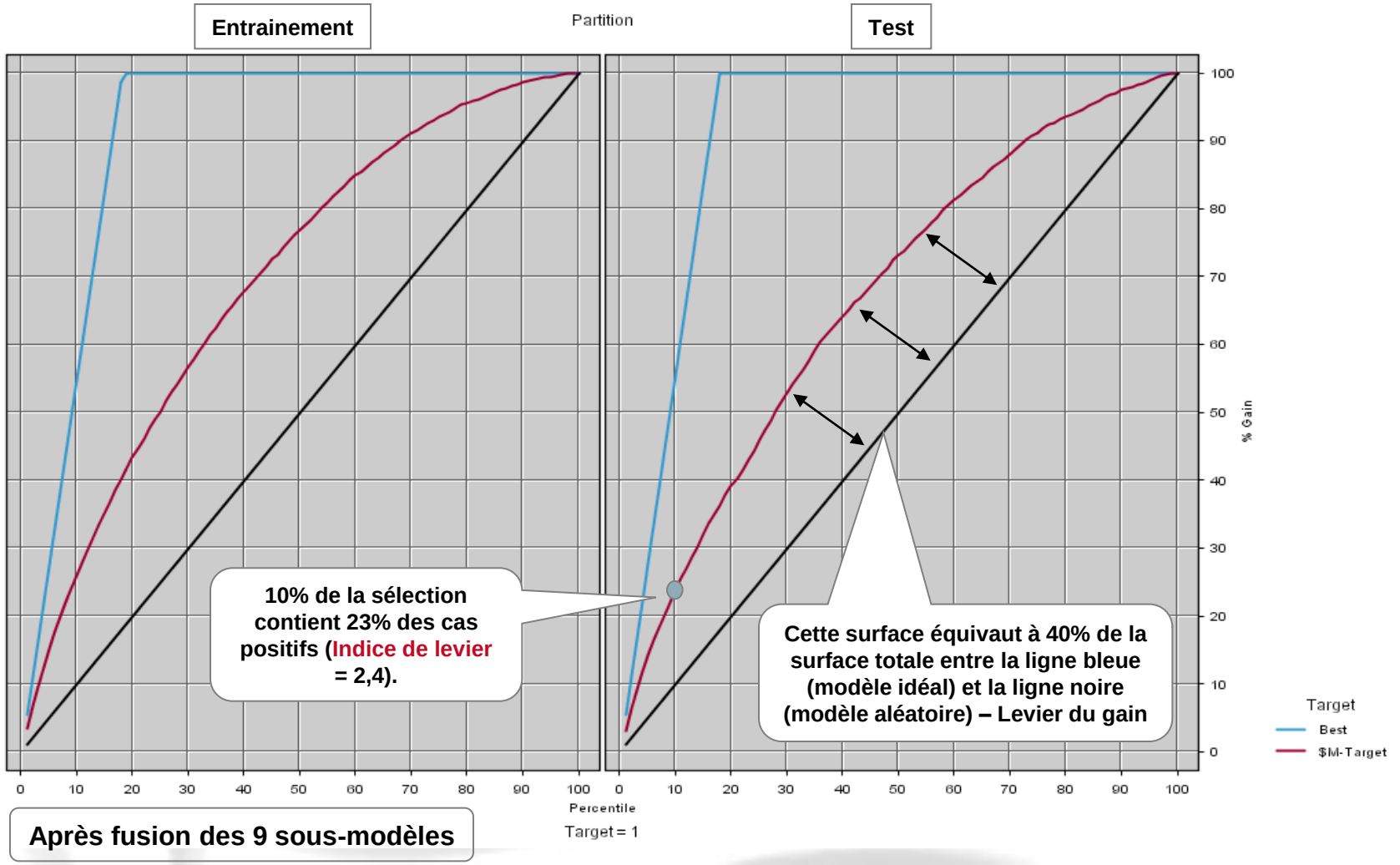
> CONFIANCE :

- La probabilité que le contribuable classé dans une "feuille" soit un cas positif (prévision du degré de réussite)
- La confirmation de cette probabilité quand le modèle est appliqué sur la population test

> CONFIRMATION AVEC LE "PLOT CHART" :

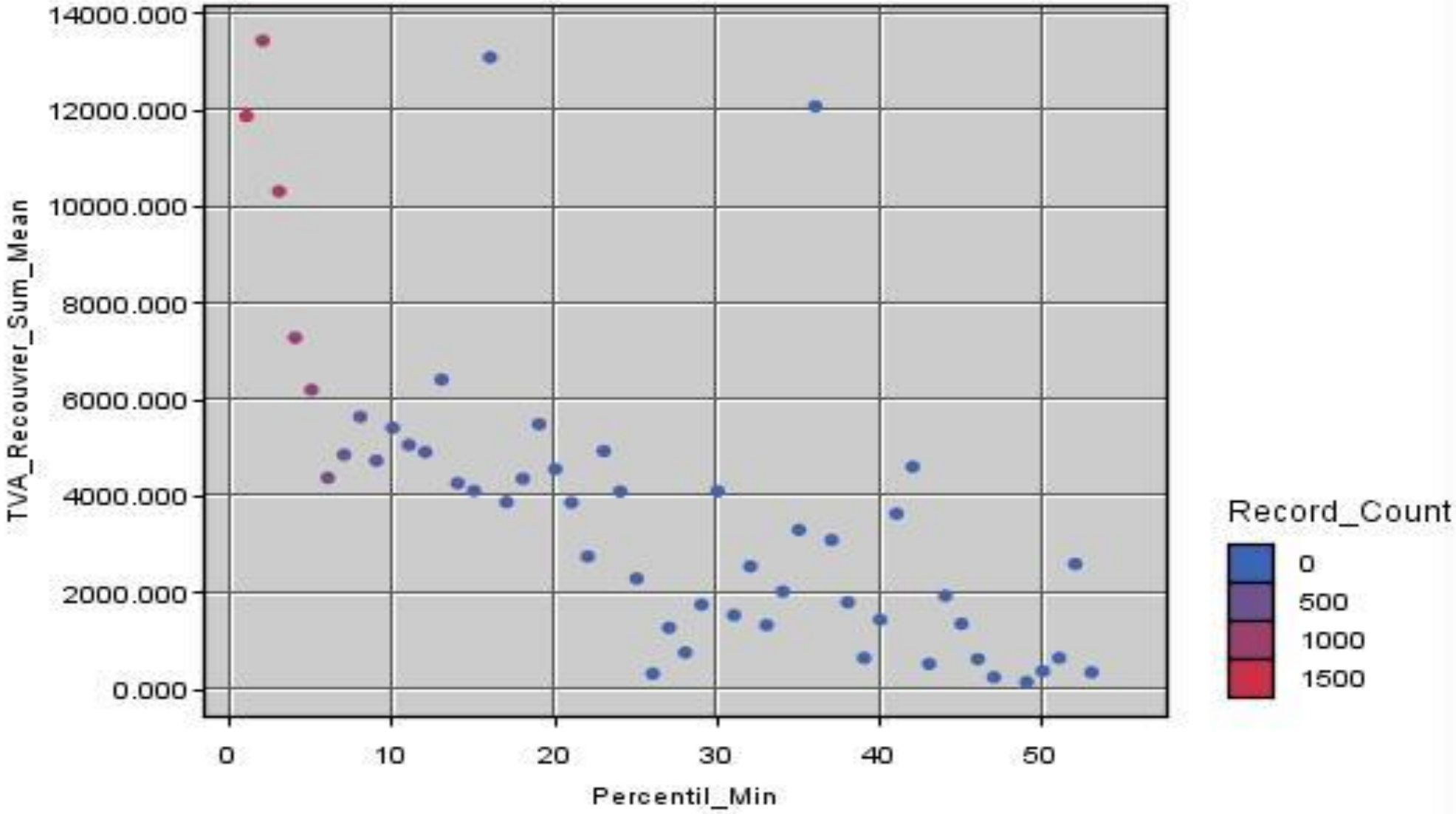
- Les résultats de contrôle vont enfin confirmer si la prévision du modèle est correcte

Levier : le gain du modèle



Après fusion des 9 sous-modèles

Exemple de confirmation par "Plots chart" DM3



Déploiement

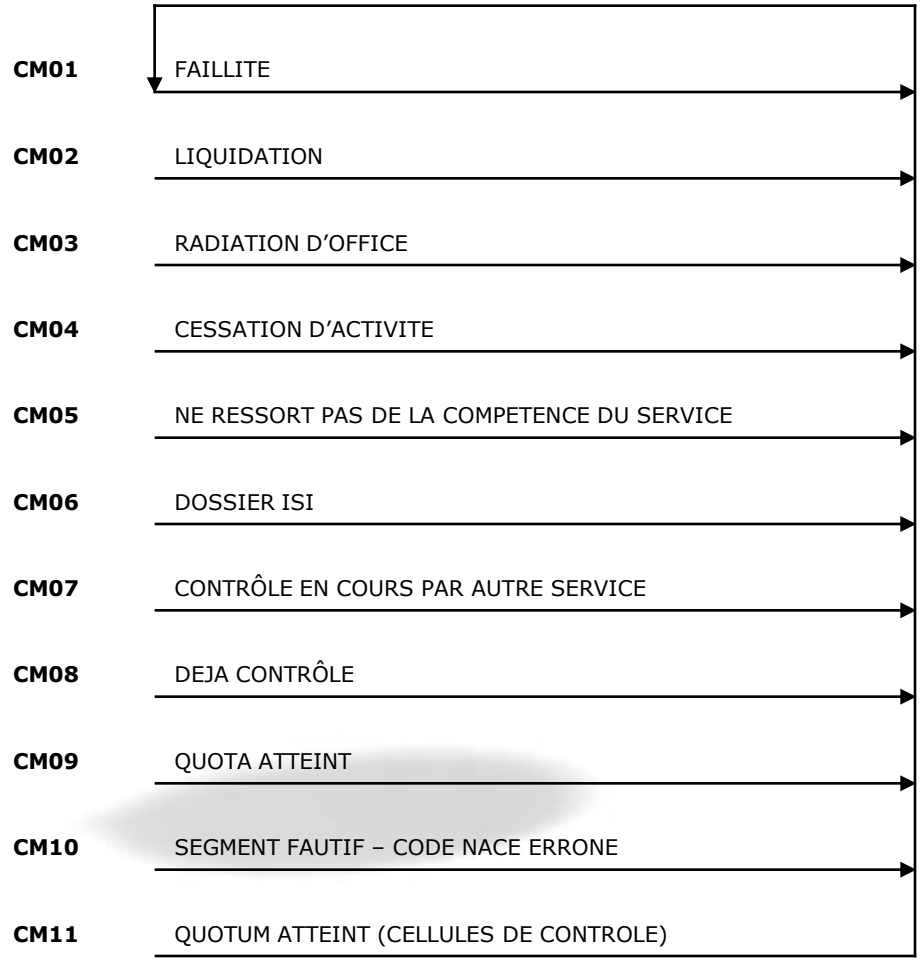
> Implémentation du modèle sur la population actuelle

> Le résultat final est présenté sous forme d'une liste de contribuables à haut risque et un chiffre qui indique le niveau de risque

> Le DM peut fournir les indicateurs de séléction de chaque dossier individuel, mais ceci ne constitue pas une relation avec une typologie de fraude connue

> Les services locaux doivent respecter le niveau de risque et ne peuvent désélectionner que pour des motifs determinés

➔



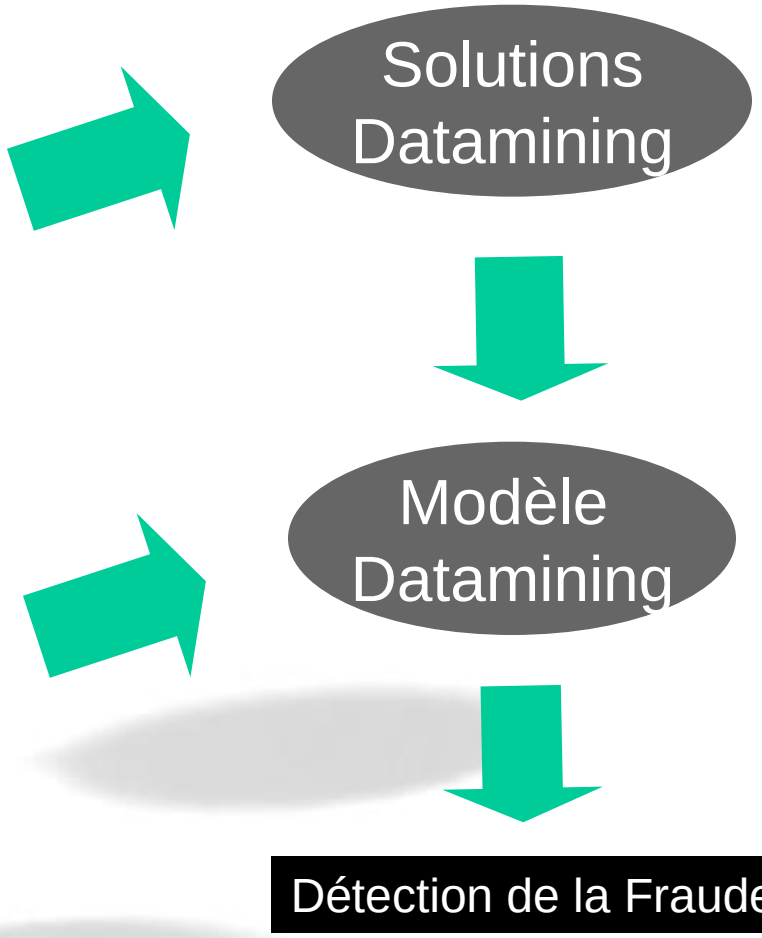
Exemple de méthodologie

Données historiques – pour l'entraînement

ID	SIC	Code Postal	Marge(N-1)/ Marge(N)	Fraude
Q870GJ	631	1060	10%	**Faux**
A980NB	436	5000	33%	**Faux**
V934VK	643	2000	99%	**Vrai**

Données actuelles – pour la production

ID	SIC	CodePo stal	Margin(N-1)/ Margin(N)	Fraude
K453LK	234	1000	100%	**?**
L634HH	235	3000	50%	**?**
V363KH	645	1020	200%	**?**
L085JV	234	2000	30%	**?**



Déploiement (exemple)

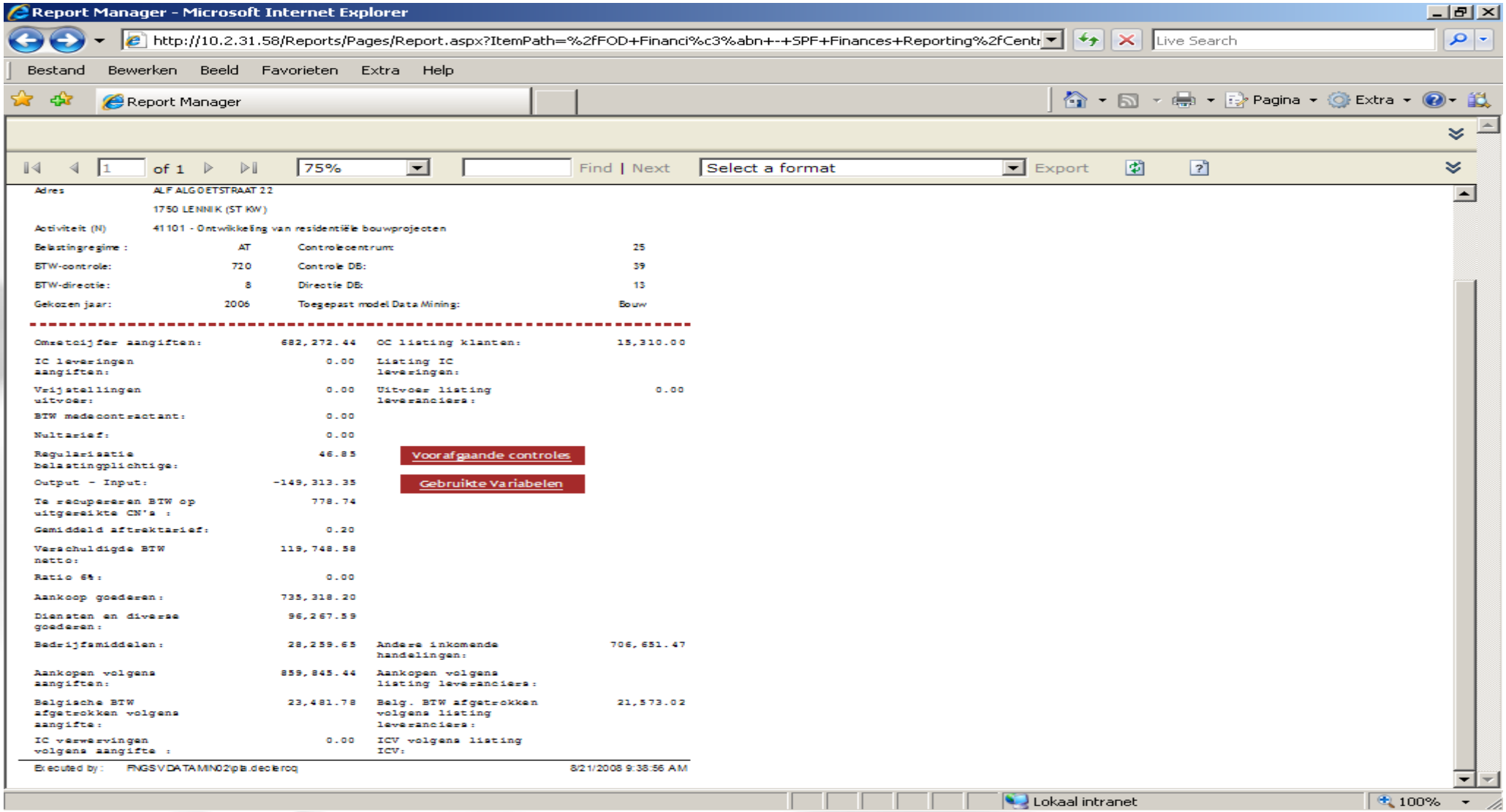
Segment	Centres de contrôle			Offices de contrôle classiques	
	NCC 2	**NCC 3**	**CC**	**TVA**	**CD**
Grandes entreprises	x	x	x		
Horeca			x		X
Mixité	x			x	
Véhicules			x	x	
Industrie			x		
Grossistes			x		
Services			x		x
Construction		x	x		
Détaillants			x		x
Faible chiffre d'affaires				x	

Rôle des sélections et du datamining au sein de l'administration

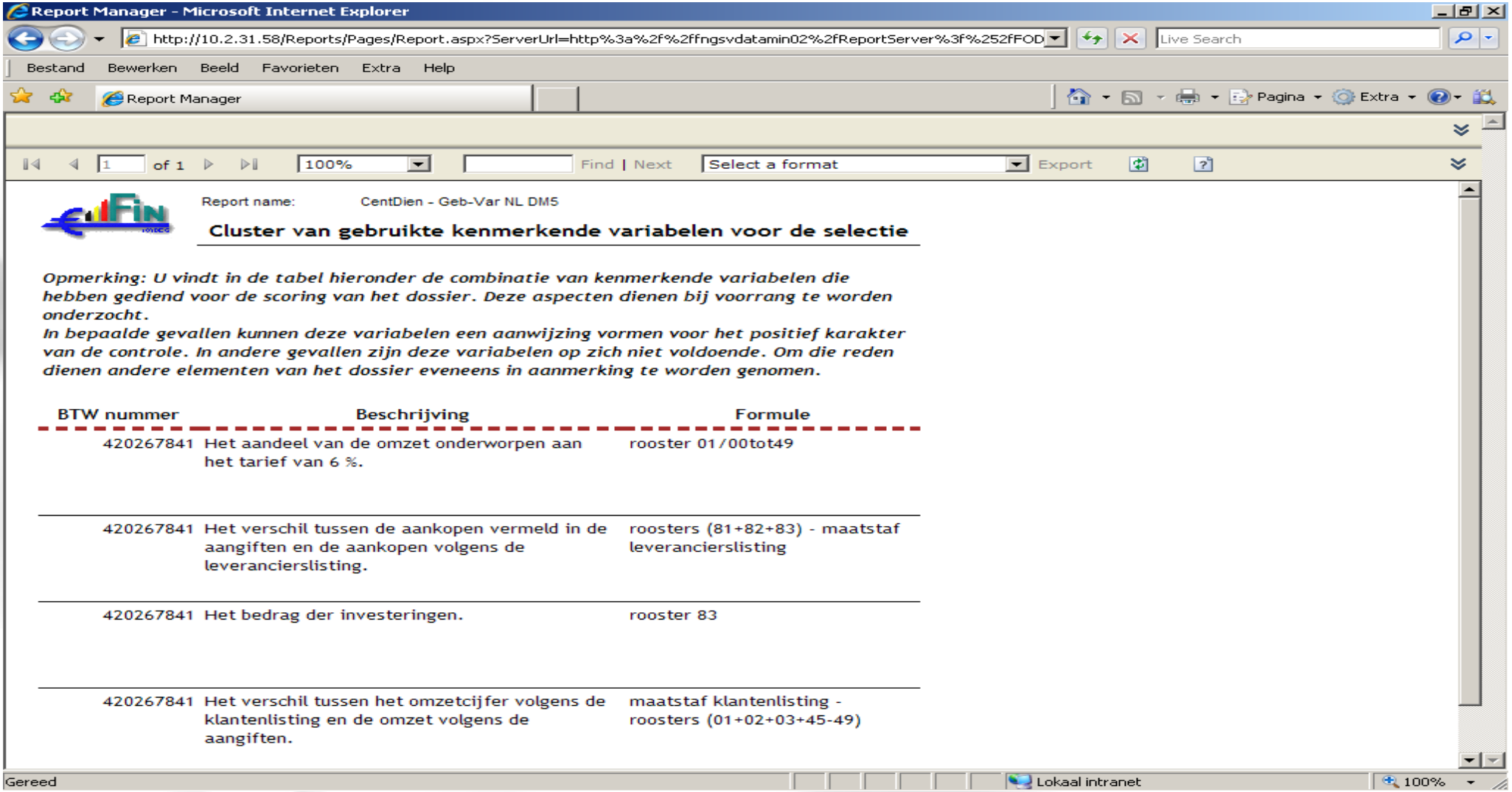
> **Les Centres de contrôle** mènent des contrôles conjoints TVA et CD
  - \+ de 40 % de leur plan de travail consiste en contrôles datamining
  - Le reste consiste en un programme fixe qui divise les contribuables en 6 groupes, dont chaque année, un groupe est sélectionné ; un programme complémentaire sélectionne les contribuables à haut risque et un autre programme consiste en contrôles ponctuels dirigés par les Services Centraux

> **Les offices de contrôle classiques** mènent des contrôles TVA ou des contrôles CD
  - 10 contrôles min par an sont sélectionnés par le datamining

Aperçu des chiffres-clés dans un fichier : fiche analytique



Cluster des différentes variables utilisées dans la sélection



Report Manager - Microsoft Internet Explorer

http://10.2.31.58/Reports/Pages/Report.aspx?ServerUrl=http%3a%2f%2ffngsvdatamin02%2fReportServer%3f%252fFOD

Live Search

Bestand Bewerken Beeld Favorieten Extra Help

Report Manager

Pagina Extra

1 of 1 100% Find | Next Select a format Export

Report name: CentDien - Geb-Var NL DM5

Cluster van gebruikte kenmerkende variabelen voor de selectie

Opmerking: U vindt in de tabel hieronder de combinatie van kenmerkende variabelen die hebben gediend voor de scoring van het dossier. Deze aspecten dienen bij voorrang te worden onderzocht.
In bepaalde gevallen kunnen deze variabelen een aanwijzing vormen voor het positief karakter van de controle. In andere gevallen zijn deze variabelen op zich niet voldoende. Om die reden dienen andere elementen van het dossier eveneens in aanmerking te worden genomen.

BTW nummer	Beschrijving	Formule
420267841	Het aandeel van de omzet onderworpen aan het tarief van 6 %.	rooster 01/00tot49
420267841	Het verschil tussen de aankopen vermeld in de aangiften en de aankopen volgens de leverancierslisting.	roosters (81+82+83) - maatstaf leverancierslisting
420267841	Het bedrag der investeringen.	rooster 83
420267841	Het verschil tussen het omzetcijfer volgens de klantenlisting en de omzet volgens de aangiften.	maatstaf klantenlisting - roosters (01+02+03+45-49)

Gereed Lokaal intranet 100%

Datamining 1 : un premier exercice

- Basé sur la technique du réseau neuronal;
- Données et analyses établies sur base trimestrielle (on examine chaque déclaration trimestrielle sur base individuelle);
- Limité aux assujettis trimestriels (289.231 personnes; 15.461 personnes sélectionnées);
- Basé sur les données de 2002 (période contrôlée 2001-2002);
- Traitement durant la période 2003-2004;
- Montant total de taxes récolté : 67.245.935 € sur 12.024 contrôles (+/- 5.000 €/controle ➔ cut off : 2.500 €)
- En comparaison avec une sélection manuelle, DM 1 a obtenu un meilleur résultat (29,70% au lieu de 26,17% des contrôles productifs sont des cas positifs (> 2.500 €))
- Les sélections manuelles libres confirment souvent les sélections datamining (pas statistiquement fondé)

Résultats du DM 1

- Les résultats du 1er DM sont les suivants (ils peuvent évidemment être améliorés) :

  - 23,36% des contrôles DM 1 peuvent être qualifiés de cas positifs (moins que l'objectif fixé de 30%)

  - les résultats d'un contrôle datamining peuvent être nuls (une méthode statistique n'est pas une garantie de succès à 100%)

- En comparaison avec une sélection manuelle, DM 1 a obtenu un meilleur résultat (29,70% au lieu de 26,17% des contrôles productifs sont des cas positifs)
- Les sélections manuelles libres confirment souvent les sélections Datamining (pas statistiquement fondé)
- Les gains moyens de DM2 sont supérieurs à ceux de DM1.

DM 2 et la fraude classique

> 9 sous-modèles

- 7 Secteurs : HORECA
  magasins de détail
  Grossistes
  garagistes (véhicules)
  Industrie
  services (ICT & consultance)
  construction
- 2 thématiques : assujettis TVA partiels (assujettis TVA avec droit à déduction partiel – art. 13 de la 6ème Directive); entreprises avec un faible chiffre d'affaires

Datamining 2

> Implique tous les assujettis : déclarations TVA mensuelles et trimestrielles

> Basé sur les données de 2003 et les données historiques : toutes les données sont analysées sur une base annuelle (données compressées)

> Basé sur la méthodologie de l'arbre de décision et diffère totalement de DM1 (seuils etc ...)

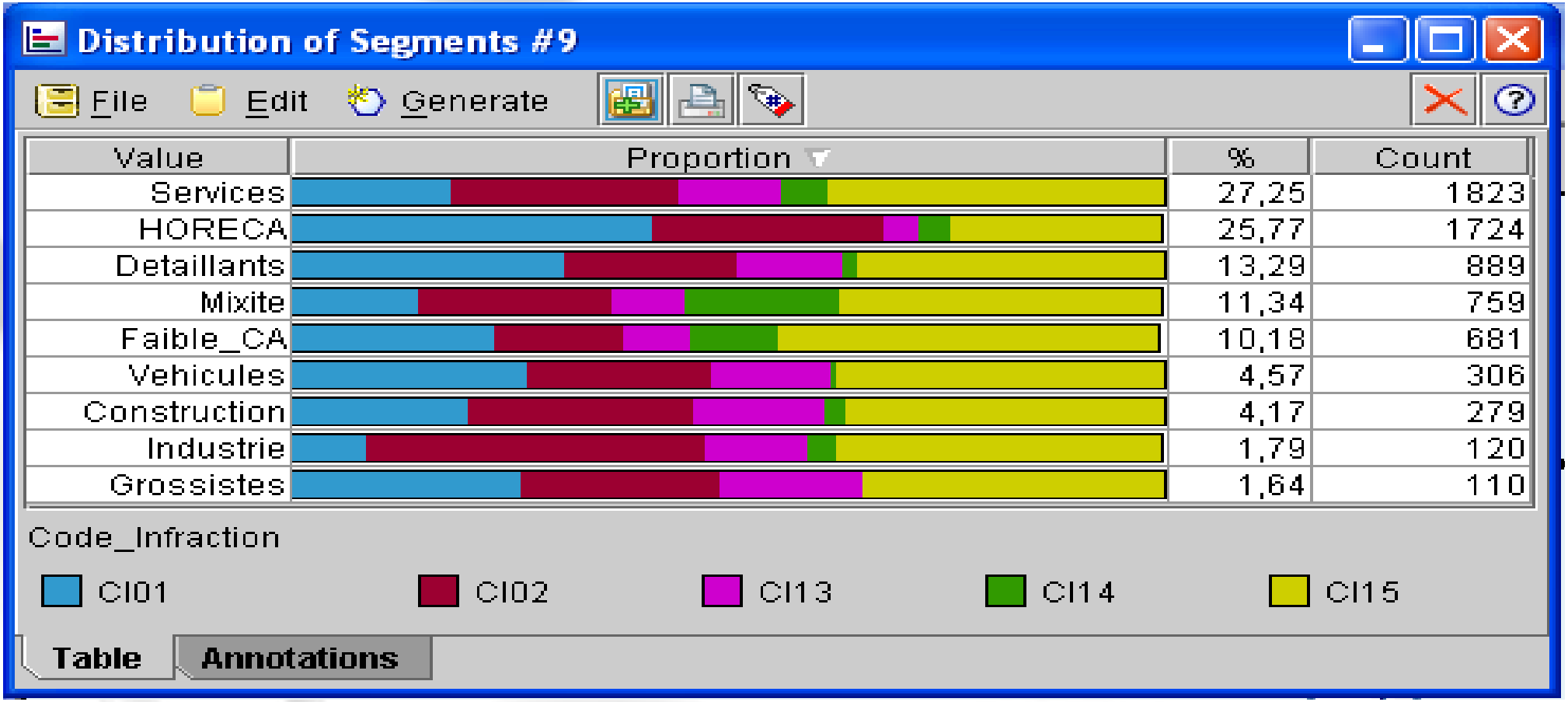
> Offre un outil de reporting pour la distribution des indicateurs de risque pris en compte pour chaque assujetti sélectionné

Datamining 2 : résultats

- 13.948 dossiers à contrôler (prévision)
- 40 % de la capacité des contrôles des CC est réservée pour les contrôles datamining
- Total TVA récolté : 47.347.839 EUR sur 6.992 contrôles
- Total taxes récolté : 210.294.756 EUR sur 14.996 contrôles
- Contrôles menés durant la période 2004-2005 sur les déclarations de la période 2003-2004

Analyse des infractions DM2

Les 5 types d'infractions les plus fréquentes dans DM2, par segment.



CI01 : CA non déclaré - CI02 : Opértions assimilées à une prestation ou une livraison à titre onéreux - CI04 : application du taux de TVA - CI13 : Déduction - CI15 : Autres déductions de TVA

Résultats DM2 pour le secteur HORECA

Sur 36.559 contribuables du segment HORECA, 4.260 ont été proposés pour le datamining.

Finalement 2.221 dossiers ont été retenus pour un contrôle.

Le montant moyen de TVA obtenu par contrôle et par année s'élève à 3.099 EUR. Cut off déterminé dans la modélisation fixé à 11.699,53 EUR

Le montant moyen d'impôts obtenu par contrôle et par année s'élève à 10.599,36 EUR

Résultats DM3

Basé sur les données de 2004

Nombre de contrôles : 12.875

Montant récolté : 204.373.880 EUR

Période contrôlée : 2004-2005

Résultats DM2 et DM3

	DM2		DM3	
	Période contrôlée 2005-2006		Période contrôlée 2006-2007	
SEGMENTS	VAT	Income Tax	VAT	Income Tax
1 - Construction	8.291.359,98	20.148.198,42	6.411.949,02	14.563.934,95
2 - Détaillants	9.433.274,57	26.594.137,19	5.897.470,87	18.601.875,82
3 – Faible chiffre d'affaires	3.414.133,63	90.983,12	1.822.431,37	17.019,45
4 – Grossistes	8.141.056,42	19.879.369,11	9.357.395,86	30.697.332,55
5 - Pubs & Restaurants	4.673.220,49	24.757.297,62	5.009.849,13	9.642.974,25
6 - Industrie	28.791.656,55	32.484.453,47	7.635.159,86	19.274.495,21
7 - Mixité	7.491.071,00	32.783,24	1.903.534,66	145.521,90
8 - Services	10.905.823,81	35.859.098,08	16.551.251,18	43.640.404,71
9 - Véhicules	6.817.733,86	4.861.493,84	9.826.366,47	3.374.912,78
TOTAL	87.959.330,31	164.707.814,09	64.415.408,4	139.958.471,6

Résultats DM4

> Pas de changements dans les modèles pour DM4

> 11.346 dossiers contrôlés

> résultat global : 207.538.689,35 EUR

> Le feedback des DM précédents a permis les premiers ajustements pour la sélection du DM5 :

- Nouveau segment pour les grandes entreprises (CA ou achats > 5.000.000 EUR)
- Application des nouveaux codes NACE introduits depuis le 1/1/2008 sur la segmentation
- Nouveaux ajustements pour les données suite à la création des UTVA

Résultats DM4

	TVA	Income Tax
1 - Construction	8.093.293 €	16.321.023 €
2 - Détaillants	8.324.960 €	31.465.568 €
3 – Faible chiffre d'affaires	1.800.523 €	2.759.330 €
4 – Grossistes	9.170.540 €	21.303.071 €
5 - Pubs & Restaurants	3.973.800 €	10.911.261 €
6 - Industrie	15.513.684 €	32.926.309 €
7 - Mixité	2.705.518 €	40.453 €
8 - Services	11.959.801 €	40.318.371 €
9 - Véhicules	2.531.187 €	2.250.294 €
TOTAL	**64.073.306 €**	**158.295.680 €**

Itérations récentes

> Agrégation des valeurs sur base annuelle

> Extension vers tous les assujettis déposants

> Résultats récents :

- DM 5 : 149.034.050 € sur 12.577 corrections
  (moyenne : 11.856 €/contrôle)
- DM 6 : 124.971.904 € sur 12.721 corrections
  (moyenne : 9.825 €/contrôle)
- DM 7 : 28.693.032 € sur 4.799 corrections
  (moyenne : 5.979 €/contrôle)

Datamining pour les opérateurs défaillants

Fraude Carrousel TVA

Approches différentes

> Fraude TVA organisée <> fraude classique

> 1 entreprise sur 5.000 <> peut-être 1 entreprise sur 2

> 8 typologies <> des dizaines de modus operandi

> Les dommages financiers sont différents

> Gravité différente

> Structures particulières <> structures économiques réelles

> Impact sur la TVA <> impact TVA & taxes directes

> Fraude caractérisée <> fraude diffuse

> Possibilité de réduction drastique / difficulté d'appréhension

> Organisée : 30% / classique : 70%

Datamining et la fraude carrousel : principes

> Modélisation basée sur différentes typologies (opérateurs défaillants, tampons, conduit companies, facturateurs croisés,…)

> Datamining est désormais l'outil courant utilisé pour la détection des nouveaux carrousels

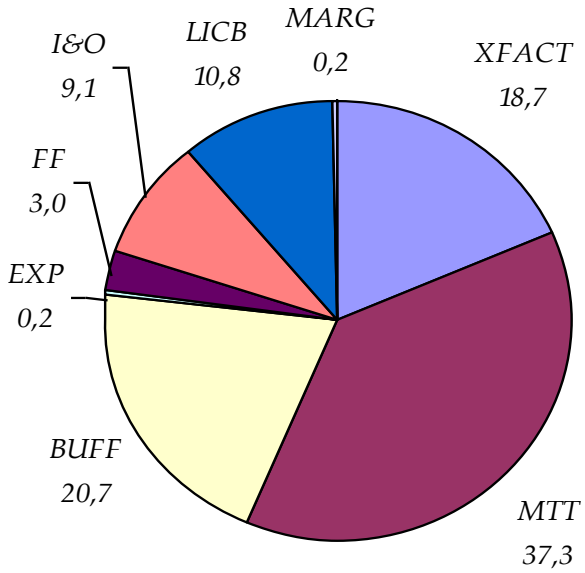
> La modélisation est basée sur 1.519 cas positifs connus du passé, mixés avec 15.000 cas TVA "normaux"

> La modélisation et la production sont appliquées sur des déclarations mensuelles et trimestrielles (asap)

> Principe général : coût de mauvaise classification = élevé (100%)

> Les cas exceptionnels rendent la modélisation et la détection plus faciles: les valeurs et écarts extrêmes (seuls les cas moins importants peuvent échapper)

Résultats de l'exploration (2001)
8 modus operandi (typologies) ont été découverts



Détection sur la base des 8 modus operandi

1. Une typologie est détectable en croisant les données (opérateur défaillant) : non-dépôt de déclaration périodique (stream spécifique pour le crosschecking des données VIES et des données déclaratives)

2. 5 typologies détectables par l'élaboration d'un profil/modèle DM

3. Autres typologies pour lesquelles la détection automatisée ne peut s'appliquer (ex: fausses factures)

Modèles exploités à l'heure actuelle

3 modèles avec exploitation quotidienne :

> Facturateurs croisés (X – invoicing)

> Buffers (tampons)

> In & Out (import / export)

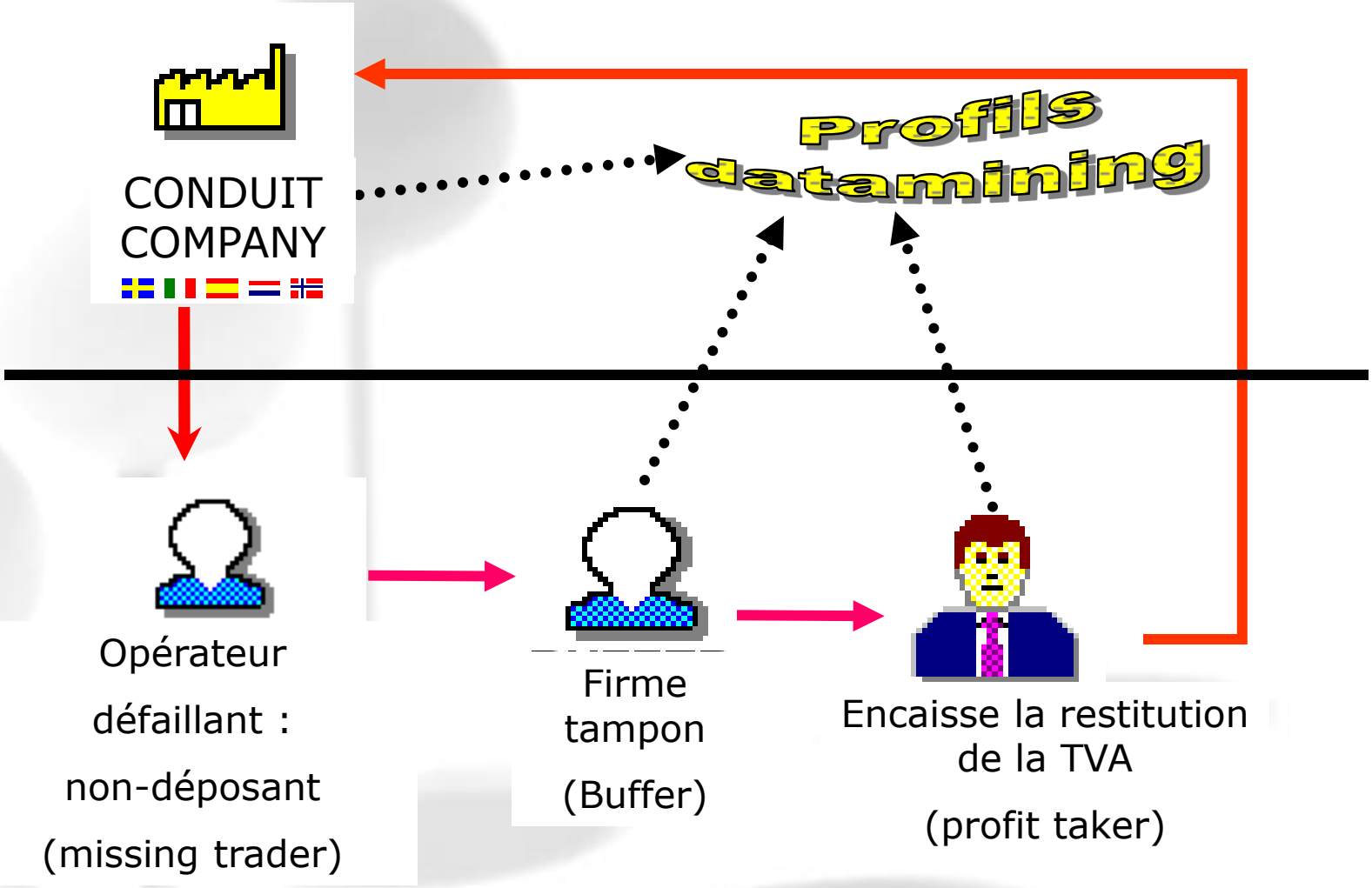
Et 3 autres modèles :

> Taxation sur la marge bénéficiaire

> « Profiteurs » (encaisseurs du crédit TVA)

> Opérateurs défaillants (missing trader sur base de données d'identité)

Schéma classique de la fraude carrousel



Firme « tampon» : profil de la déclaration

2000(12)

00 :	0,00	71 :	0,00
01 :	0,00	72 :	821,64
02 :	0,00	81 :	206.171,66
03 :	206.253,78	82 :	4.390,69
45 :	0,00	83 :	1.267,08
46 :	0,00	84 :	0,00
47 :	0,00	85 :	0,00
48 :	0,00	86 :	0,00
49 :	0,00	87 :	0,00
54 :	43.313,30	88 :	0,00
55 :	0,00	91 :	
56 :	0,00		
57 :	0,00		
58 :	0,00		
59 :	44.134,94		
61 :	0,00		
62 :	0,00		
63 :	0,00		
64 :	0,00		
00/49 :	206.253,78		

Conduit companies : profil de la déclaration

869.915.004
2005(12)

00:	0,00	71:	0,00
01:	0,00	72:	4.763,46
02:	0,00	81:	19.262.979,71
03:	0,00	82:	66.024,85
45:	0,00	83:	989,42
46:	16.926.777,00	84:	177.097,50
47:	2.619.388,91	85:	4.461,92
48:	0,00	86:	16.668.920,22
49:	0,00	87:	118,40
54:	0,00	88:	0,00
55:	3.331.110,42	91:	
56:	14,37		
57:	0,00		
58:	0,00		
59:	3.335.523,99		
61:	37.323,13		
62:	37.694,35		
63:	6,96		
64:	0,00		

00/49: 19.546.165,91

Comparaison fraude >< pas de fraude

Profil frauduleux

401.785.183
2000 (09)

00:	0,00	71:	3.015,63
01:	0,00	72:	824,25
02:	0,00	81:	44.310.023,78
03:	21.282.918,38	82:	61.815,12
45:	0,00	83:	0,00
46:	23.248.674,72	84:	0,00
47:	0,00	85:	0,00
48:	0,00	86:	23.099.355,60
49:	0,00	87:	0,00
54:	4.469.412,87	88:	0,00
55:	4.850.864,68	91:	
56:	0,00		
57:	0,00		
58:	0,00		
59:	9.318.086,16		
61:	0,00		
62:	0,00		
63:	0,00		
64:	0,00		

00/49: 44.531.593,10

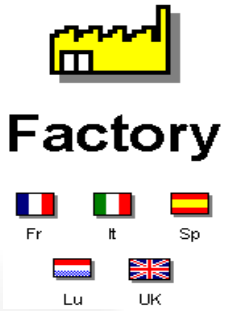
Profil non frauduleux

435.305.514
032008E*

00:	17.230,85	71:	0,00
01:	9.856,54	72:	1.338.157,09
02:	0,00	81:	38.171.298,91
03:	5.965.086,57	82:	9.265.036,20
45:	3.722.269,30	83:	1.273.110,96
46:	0,00	84:	161.075,46
47:	31.932.368,99	85:	7.734.236,89
48:	0,00	86:	5.255.489,13
49:	1.371.786,89	87:	16.146.892,87
54:	1.253.259,60	88:	0,00
55:	446.326,71	91:	
56:	3.613.517,32		
57:	0,00		
58:	0,00		
59:	7.233.528,48		
61:	0,00		
62:	0,00	IND :	21.04.2008
63:	587.753,37	***	001 ***
64:	5.485,61		

00/49: 40.275.025,36

FACTURATION CROISEE

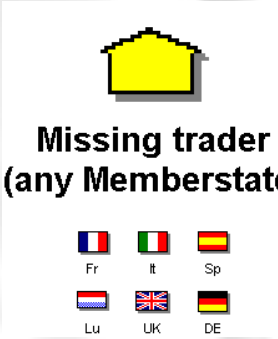


Réel achat intracommunautaire

Livraison en Belgique fictive avec facturation de la TVA

Crossed invoicer

Vente intracommunautaire fictive sans TVA



Vente réelle avec TVA

Facturation croisée : cas historique et cas détecté

Cas historique

401.785.183
2000 (09)

00:	0,00	71:	3.015,63
01:	0,00	72:	824,25
02:	0,00	81:	44.310.023,78
03:	21.282.918,38	82:	61.815,12
45:	0,00	83:	0,00
46:	23.248.674,72	84:	0,00
47:	0,00	85:	0,00
48:	0,00	86:	23.099.355,60
49:	0,00	87:	0,00
54:	4.469.412,87	88:	0,00
55:	4.850.864,68	91:	
56:	0,00		
57:	0,00		
58:	0,00		
59:	9.318.086,16		
61:	0,00		
62:	0,00		
63:	0,00		
64:	0,00		

00/49: 44.531.593,10

Cas détecté par DM

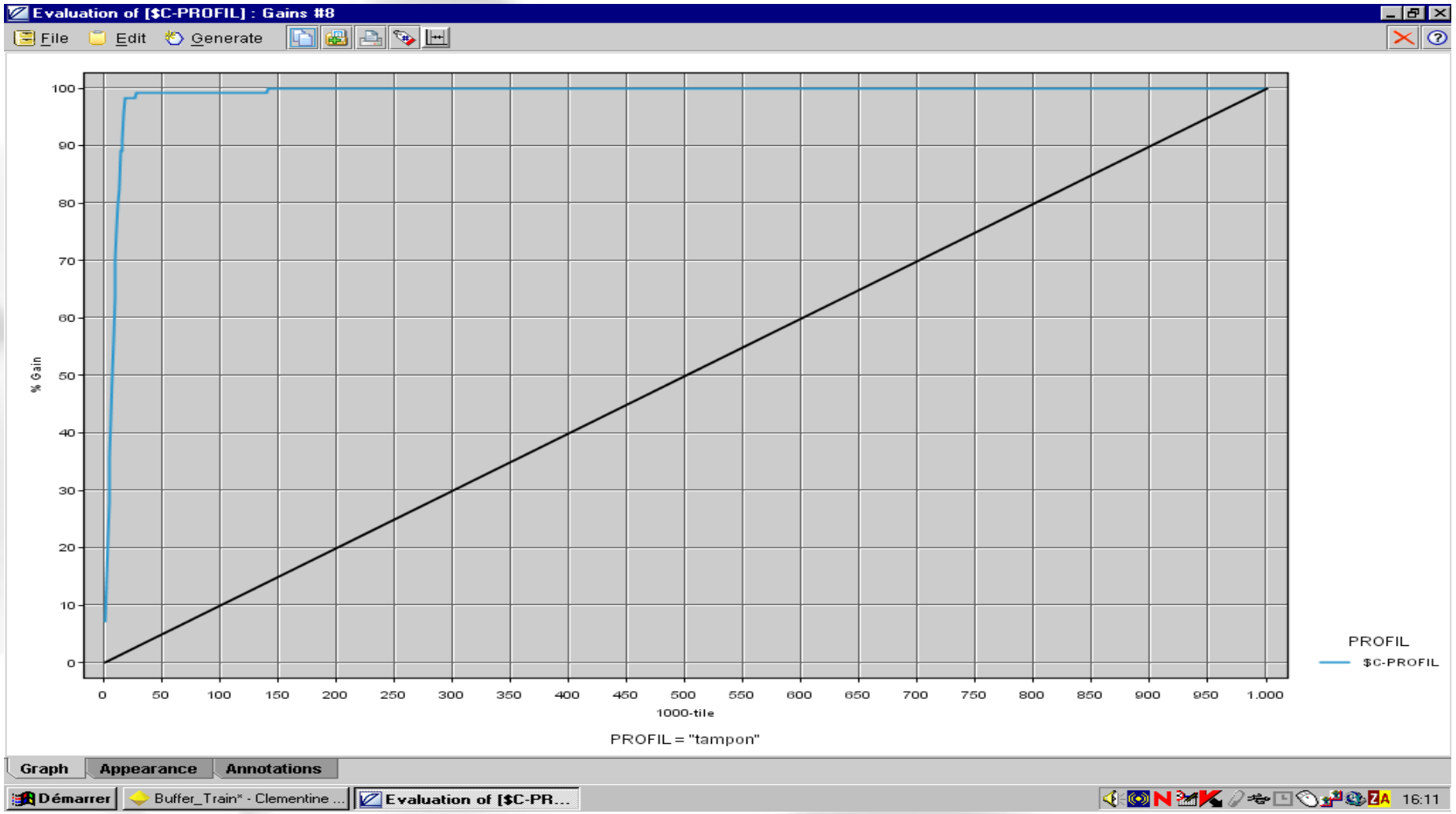
451.505.405
022008E

00:	0,00	71:	77.466,28
01:	0,00	72:	0,00
02:	0,00	81:	2.896.567,86
03:	1.240.870,00	82:	7.758,08
45:	0,00	83:	0,00
46:	1.464.500,00	84:	0,00
47:	500,00	85:	0,00
48:	0,00	86:	2.032.343,00
49:	0,00	87:	0,00
54:	260.582,70	88:	0,00
55:	426.792,03	91:	
56:	0,00		
57:	0,00		
58:	0,00		
59:	609.908,45		
61:	0,00		
62:	0,00		
63:	0,00		
64:	0,00		

IND : 19.03.2008
*** 001 ***

00/49: 2.705.870,00

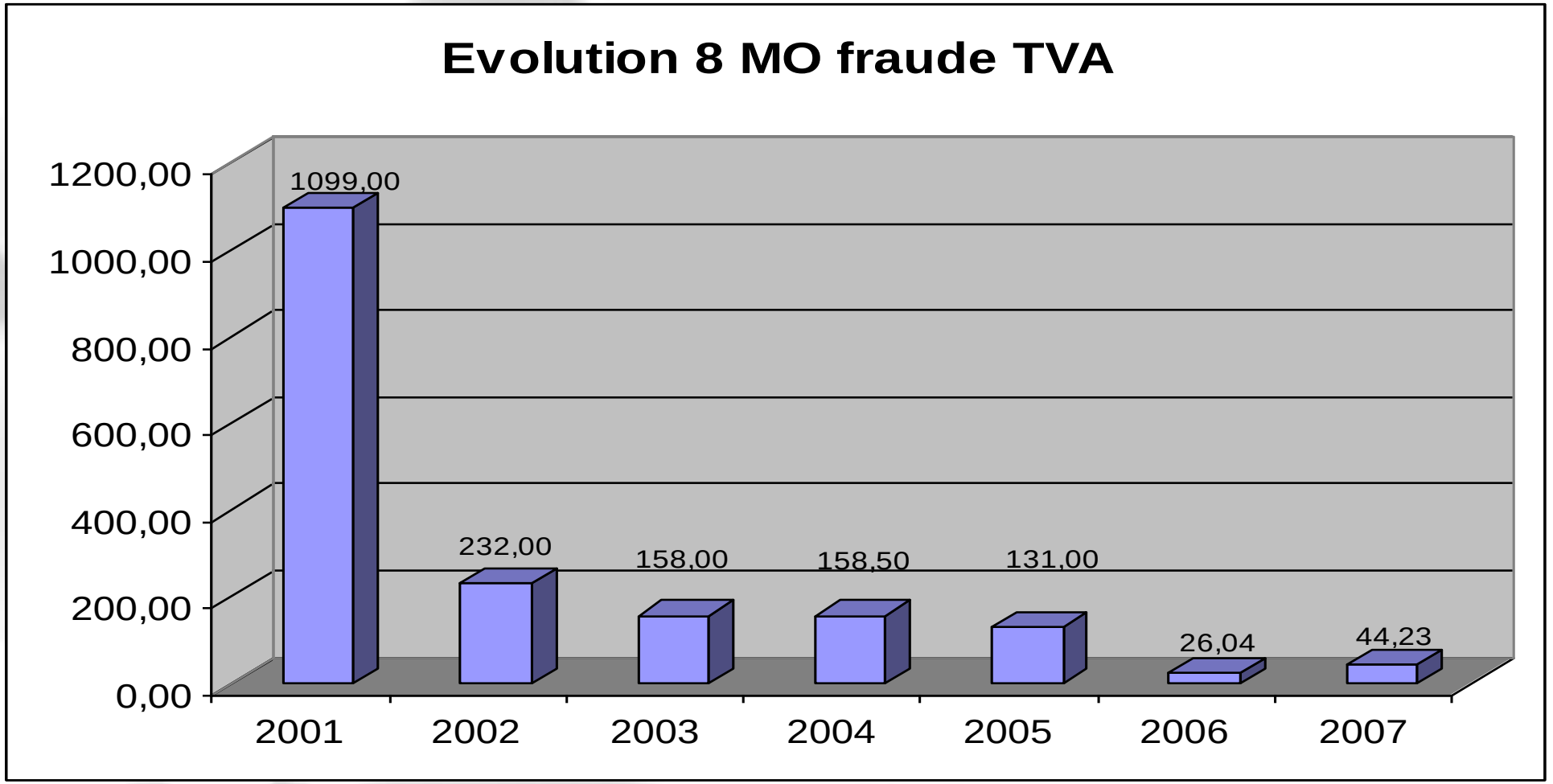
Le levier en cas de DM pour fraude carrousel



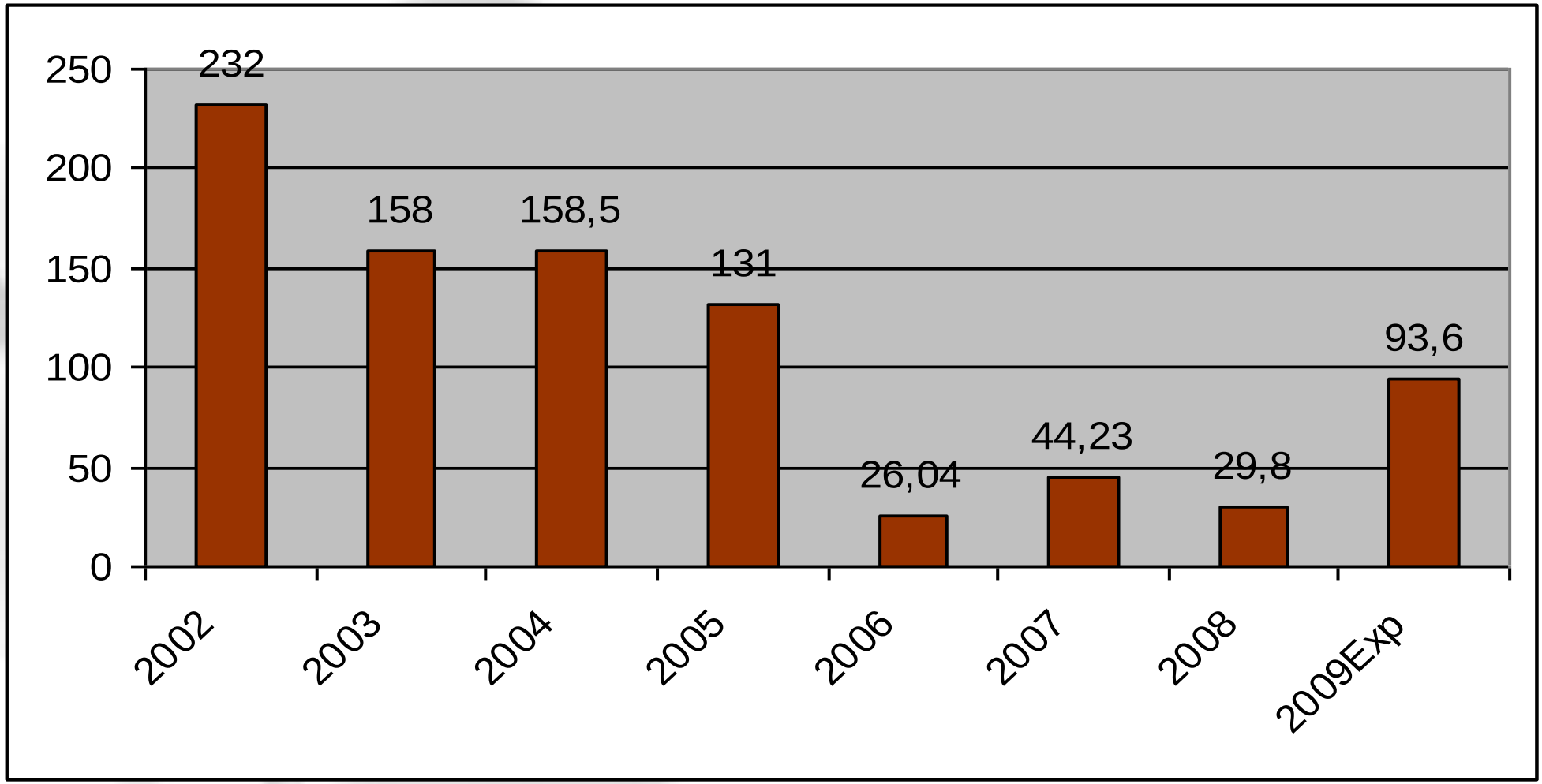
Appréciation quantifiée

		Total des déclarations TVA scannées & contrôlées	Résultat : nombre de cas positifs	1/10000e	Nom du modèle
IN&OUT	Run 1	**330694**	**14**	**0,423352102**	**IBIZA**
	Run 2	**600051**	**32**	**0,533288004**	
XFACT	Run 1	**330694**	**64**	**1,935323895**	**VALLETA**
	Run 2	**600051**	**101**	**1,683190262**	
TAMPON	Run 1	**330694**	**43**	**1,300295742**	**PARIS**
	Run 2	**600051**	**81**	**1,34988526**	

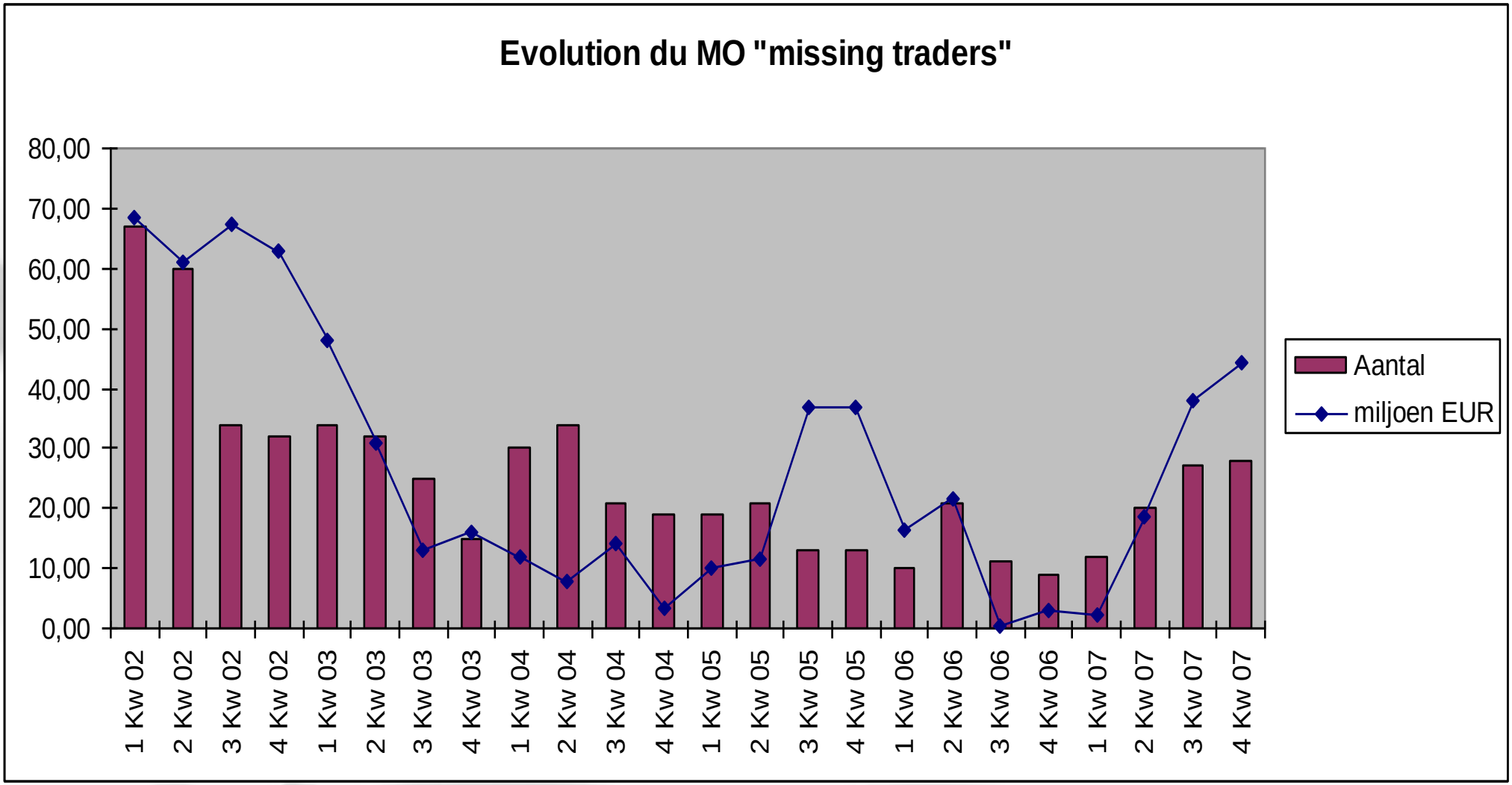
Evolution de l'impact budgétaire de la fraude carrousel : 2001 (sans DM) ➔ période avec DM



Mais toujours … vigilance !!!



Evolution de nombre de cas de fraude carrousel



Production des modèles

Données

>3 millions de déclarations TVA

>3 millions de clients sur les listes IC

>14 millions de clients belges sur les listings nationaux

DM pour carrousels : conclusions

> Les pertes de budget causées par la fraude des opérateurs défaillants TVA a chuté de 1.100 millions € en 2001 à 26,04 millions € en 2006 et 93,6 millions € en 2009.

> La population **entière** est systématiquement «scannée» par le logiciel

> Réaction rapide : 99,9% du processus de détection est fait de manière automatique dès le dépôt de la première déclaration suspecte

> La fraude se prête bien au profiling (la caractérisation est très précise !)

> Les modèles sont améliorés sans cesse en fonction des nouveaux profils de fraude

> Angle international (I & O, chaînes de fraudes internationales,… ➔ nécessitant la coopération internationale)