Statistics on Trade Data for Anti-fraud Applications
Overview of problems, patterns and examples of results obtained

Spikes, price outliers, fair prices & positive associations in 2 way tables

Domenico Perrotta

EC Joint Research Centre

Institute for the Protection and Security of the Citizen

Global Security and Crisis Management Unit

Sofia, October 31, 2013

EU's financial interests under threat:

new approaches in assessing the risk from public procurement and EU funds fraud

Approach in addressing fraud problems

1. Fraud control problem statement
2. Statistical pattern to detect in data
3. Statistical method to (develop and) apply
4. Implementation of Method
5. Presentation/web publication of results for access by data owners / authorized user
6. Fine-tune, parametrize, make method available for use to wider community of users

Addressing fraud problems
Statement of 6 hard fraud control problems

1. Formation of stocks before EU Enlargements (stockpiling).
2. Export refunds paid for goods that are not consumed in the country of export, not of appropriate quality, etc.
3. Evasion of payment of import duties.
4. Deflection of trade: imports into the EU bypassing import or antidumping duties or quotas in imports by false declaration of origin or product type.
5. Carousel: export of goods from a MS and subsequent re-import to benefit from export refunds.
6. Trade Based Money Laundering.

Addressing fraud problems

7. Anti-dumping duties
8. VAT fraud
9. …

N.B. List of problems appears to be endless

Approach in addressing fraud problems
Statistical patterns in trade data

1. *Upward (downward) spikes* in trade flows: sudden, "unexpected", unprecedented increases (decreases) in trade flows at a point in time.
2. *Price outliers:* trade flows with unit price significantly different than other comparable unit prices for some point in time or transaction.
3. *Systematic underpricing:* unit price reported in some combinations of trade flows (defined by concomitant variables) lower than other comparable flows.
4. *Positive (systematic) associations in 2- way tables*: some categories of a categorical variable have an affinity for some categories of another variable.

N.B. Relatively few patterns

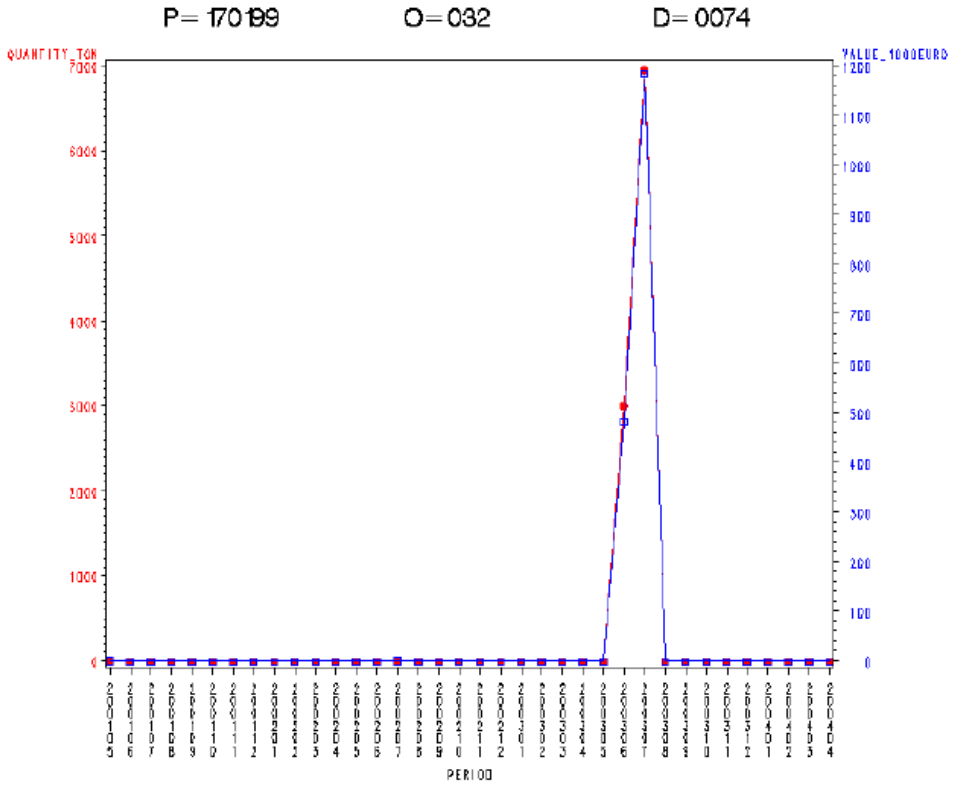
European Commission

Approach in addressing fraud problems
Matrix of Fraud Control Problems and Patterns

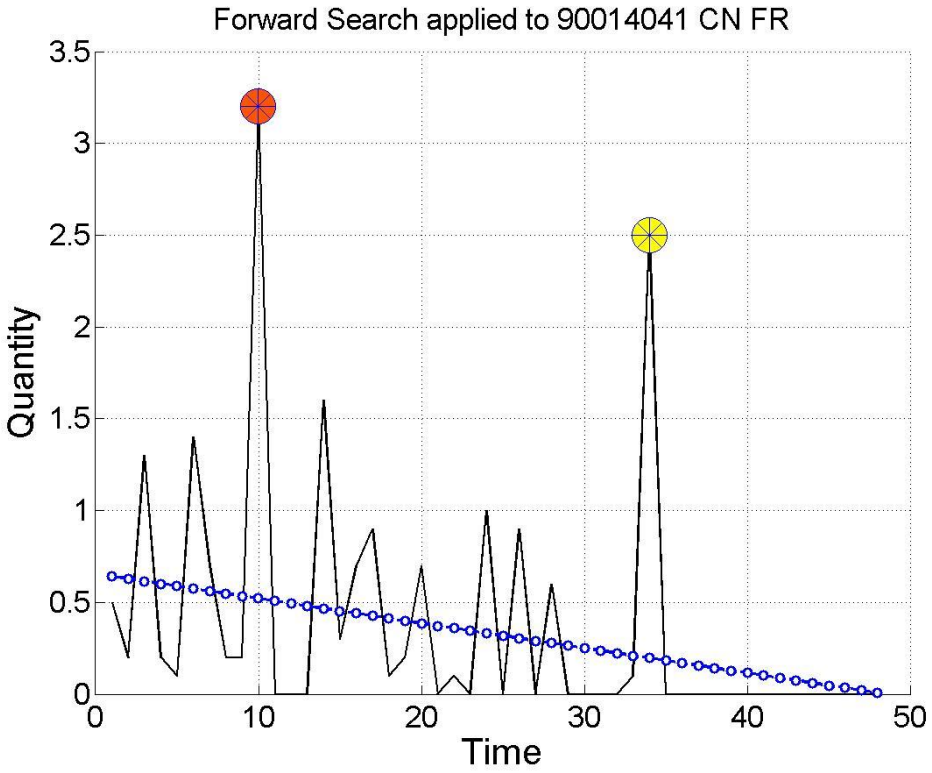
	U-Spikes	Outliers	D-Spikes	Systematic Underpricing	Systematic Associations in 2 way tables
Stockpiling	•				
Fraud in Export Refunds	•				
Evasion of import duties		• LP outliers		•	•
Deflection of Trade	•, partly		•, partly		
TBML		• HP, LP outliers			•
VAT fraud		*• LP outliers*			*•*

Pattern 1, Spikes in trade data

Example 1.1, …170199: cane or beet sugar …, from 032:Finland to 0074:Moldova

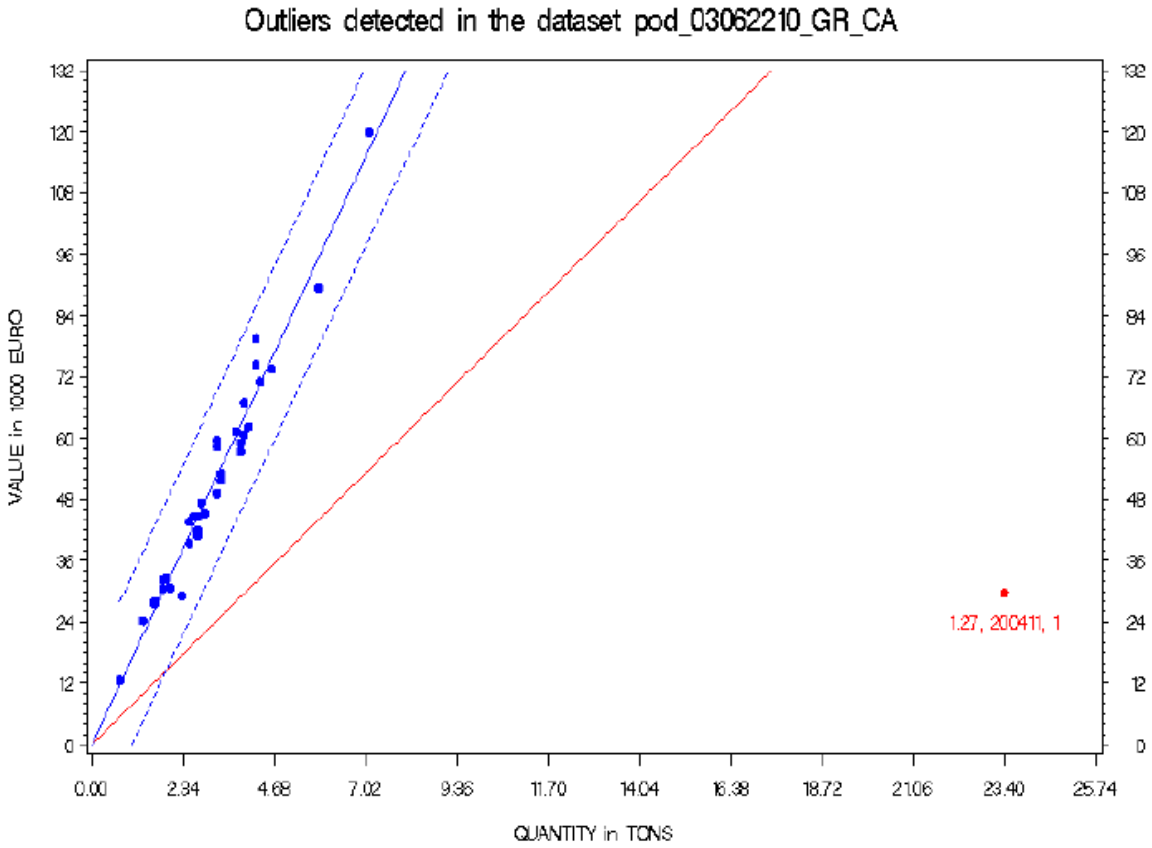


Pattern 1, Spikes in trade data



Pattern 2, Price outliers in COMEXT data

Example 2.1, Lobsters imported into Greece in 200411 at 1.27 €/Kg; average price for others 16.41 €/Kg



Detecting price outliers in MS customs declarations

Summary on outliers detected, all flows, all data owners

		Transactions Initially uploaded	Tranasactions searched for outliers	Fraction of transactions searched	LP outliers detected	Fraction of LP outliers detected	HP outliers detected	Fraction of HP outliers detected
Imports	AT	1,790,643	1,729,204	0.966	5,564	0.003	12,684	0.007
	BE	2,627,438	2,575,154	0.980	8,304	0.003	18,071	0.007
	NL	3,635,125	3,542,010	0.974	9,077	0.003	22,725	0.006
Exports	AT	2,184,677	2,087,714	0.956	6,286	0.003	13,899	0.007
	BE	3,119,601	2,856,094	0.916	5,727	0.002	13,901	0.005
	NL	3,937,934	3,726,727	0.946	7,705	0.002	19,721	0.005

Still, a large number of price outliers are detected:

outliers have to be prioritized

Detecting price outliers in MS customs declarations need to prioritize

- 1st approach, through estimated undervaluation
- 2nd approach, through detecting affinity (positive association) of importers with LP declarations

European Commission

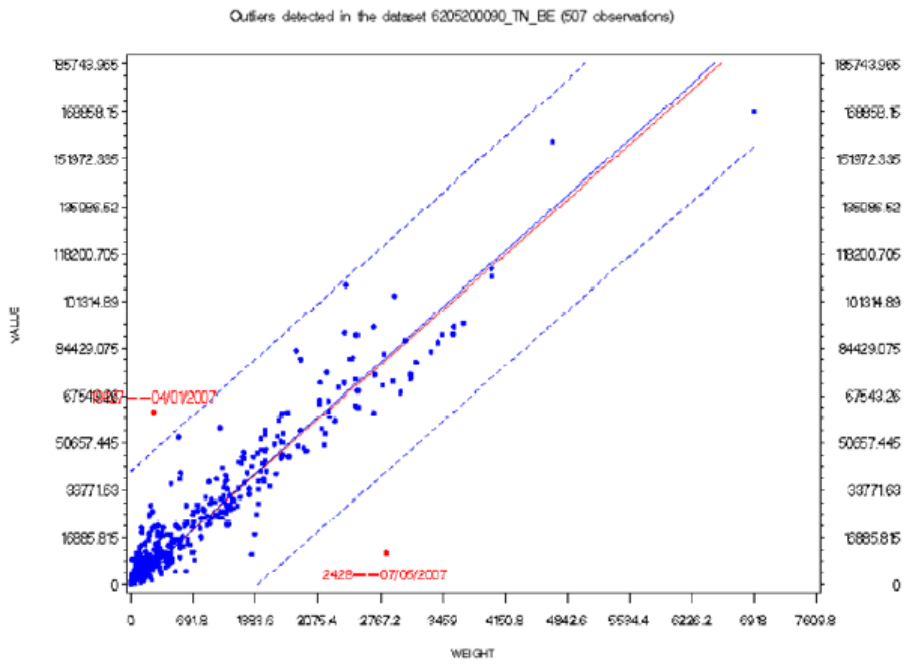
Feedback received from BE, on 1st approach ... Hmm!

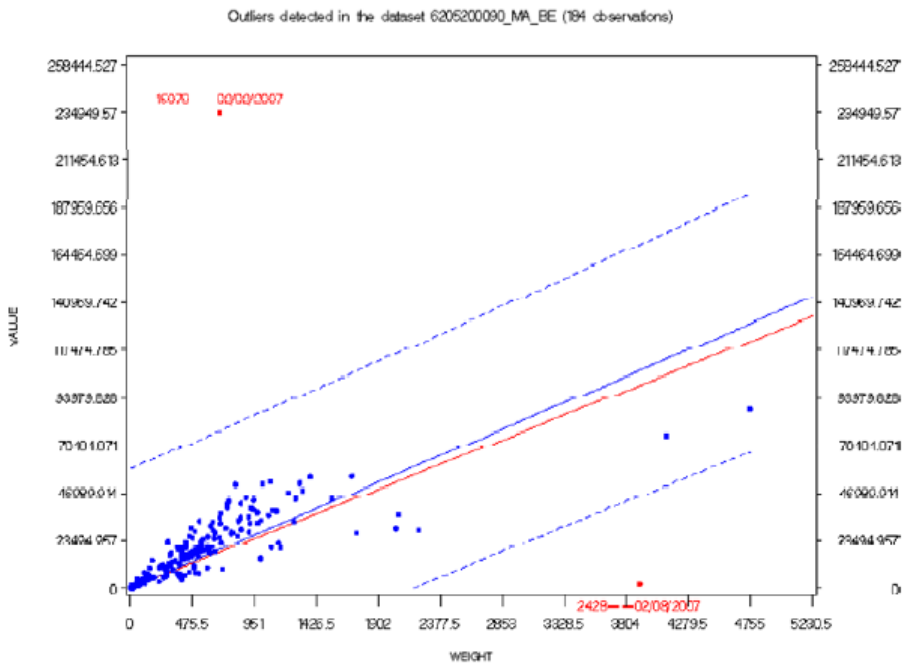
Email of	Source of error for the outlier	Finding of Data owner	P	O	D
09.05.12	Erroneous entry of product code Typing error in weight	Product code should be 0000000000 Weight should be 1000 times smaller	6203423100	TR	BE
09.05.12	Typing error in weight	Weight should be 1000 times smaller	9018311000	US	BE
09.05.14,	As in 1		8512200000	US	BE
09.05.14	Typing error in weight	Weight should be 1000 times smaller	9018908500	US	BE
09.05.15	Erroneous entry of product code	Product code should be 0000000000 Globalization declaration IMZ procedure 4271	8443321090	QU	BE
09.05.18	typing error of weight	Weight should be 1000 times smaller	9503004100	CN	BE
09.05.25	Erroneous entry of product code (globalization)	Product code should be 0000000000	8482101000	US	BE
			8708999790	BR	BE
09.05.25	typing error of weight	?	8512909000	JP	BE
			7102390000	IN	BE

European Commission

2nd prioritization: detecting positive association of companies with LP, HP declarations

Example. In BE imports, company *2428* appears in two import declarations both detected as low price outliers, in two different scatter plots of 184 and 507 imports of product 6205200090





2nd prioritization: detecting positive association of companies with LP, HP declarations

Assumption:

Data quality problems should be spread across all declarations not on specific importers.

Procedure:

The set of 2,575,154 (BE) imports explored were declared by 31,473 companies.

The 26,383 high and low price outliers detected in BE imports were declared by 4,165 companies.

Focus on these 4,165 companies involved in high or low price declarations which account for 2,138,939 declarations.

2nd prioritization: detecting positive association of companies with LP, HP declarations

Assumption:

Data quality problems should be spread across all declarations not on specific importers.

Procedure:

The set of 2,575,154 (BE) imports explored were declared by 31,473 companies.

The 26,383 high and low price outliers detected in BE imports were declared by 4,165 companies.

Focus on these 4,165 companies involved in high or low price declarations which account for 2,138,939 declarations.

2nd prioritization: detecting positive association of companies with LP, HP declarations

Assumption:

Data quality problems should be spread across all declarations not on specific importers.

Procedure:

The set of 2,575,154 (BE) imports explored were declared by 31,473 companies.

The 26,383 high and low price outliers detected in BE imports were declared by 4,165 companies.

Focus on these 4,165 companies involved in high or low price declarations which account for 2,138,939 declarations.

2nd prioritization: detecting positive association of companies with LP, HP declarations

Procedure:

Want to detect how these 4,165 companies are "positively associated" with declarations FP, LP or HP: $P_{i,j} > P_{i,+} \cdot P_{+,j}$

1. Select cells where most data lie, to reduce the number of comparisons: i.e. suppress cells of a small number of counts, typically 1 or 2 counts.
2. For each of the selected cells, test the cell for having positive mutual information by applying the *Fisher exact test* on the collapsed 2 by 2 table. Bonferroni correct for the *multiple tests* done (nominal alpha / # of comparisons).

2nd prioritization: detecting positive association of companies with LP, HP declarations

Results on Belgian imports

COUNTXY	SF	FS	NCOMP0	NCOMPA	NSIGNALS	NCATX0	NCATXA	NCATXS	NCATY0	NCATYA	NCATYS
2138939	1.00	0.37	9277	6469	358	4165	3996	339	3	3	3

i.e.

of 4,185 importers, only 339 have an affinity to LP, HP or FP outliers.

Focus on 108 of those with an affinity to LP outliers.

2nd prioritization: detecting positive association of companies with LP, HP declarations

Results on Belgian imports

	COMPANY_ID	LP	CX_S	CX_A	CX_0	FLP
1	***132ad8f05281df6c9bcd4b221a035ca9***	***2***	***2***	***2***	***2***	***1.000***
	147eeb9b9b973135103c0cce7261a9a2	2	2	2	2	1.000
	bfe265c3591bce8f4d5bbcec8d4b2f43	2	2	2	2	1.000
	610c415536760040d4939a4926b8cf82	3	3	3	4	0.750
	d56f1c4dde8bd867d3917bb0569f1083	3	3	3	4	0.750
	3af802ddae45da79a5701f141c3ef28a	5	5	9	9	0.556
…						
	d2eda8d37a49e71fbde7c246968bcfa3	54	54	6031	6031	0.009
89	baa53bfc0f88d958dccf000ccff0b36f	149	149	18306	18306	0.008
		2044	786337			
		7195		2136131		

Summary of operational results from BE customs

Of 21 investigations that have been launched by BE Customs:

On one importer false alarm -- used mechanical engineering equipment imported

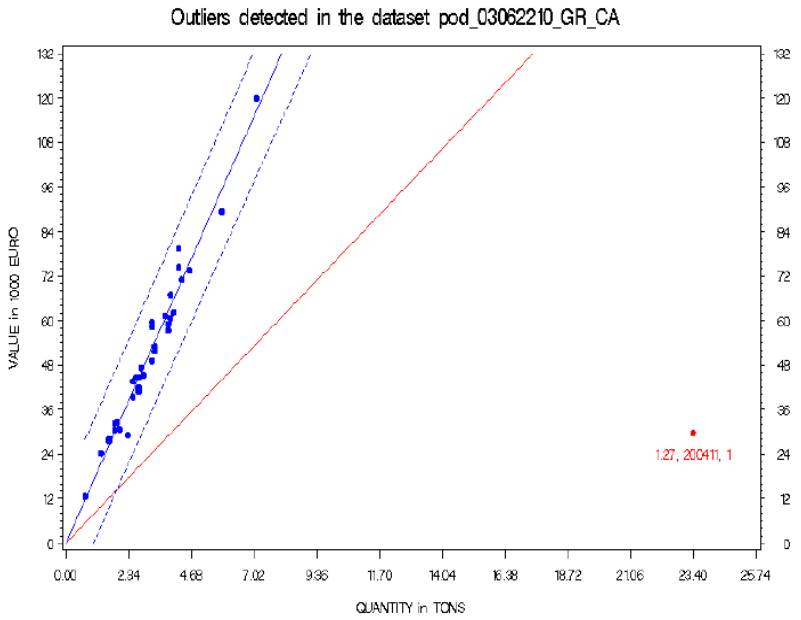
In remaining 20 investigations, importers have committed serious economic crimes:

- fraudulent bankruptcies and VAT fraud, (2 operators, 736K€ and 7.7M€);
- transfer pricing (168K€);
- TBML and customs fraud; Customs fraud and/or VAT fraud (4 importers) and 12 importers under customs services investigations.

Fair prices: a spinoff of price outliers in COMEXT data

Recall the price outlier pattern …

the estimated slope can be used as an indicator for the fair price



Fair prices: a spinoff of price outliers in COMEXT data

Estimated Fair Prices - Thirteenth COMEXT download (01/07/2007 - 30/06/2010), extracted in October 2010

Clear all Filters | Filtered for: **'Product'** starts with **'03062210'** and **'Origin'** starts with **'CA'**

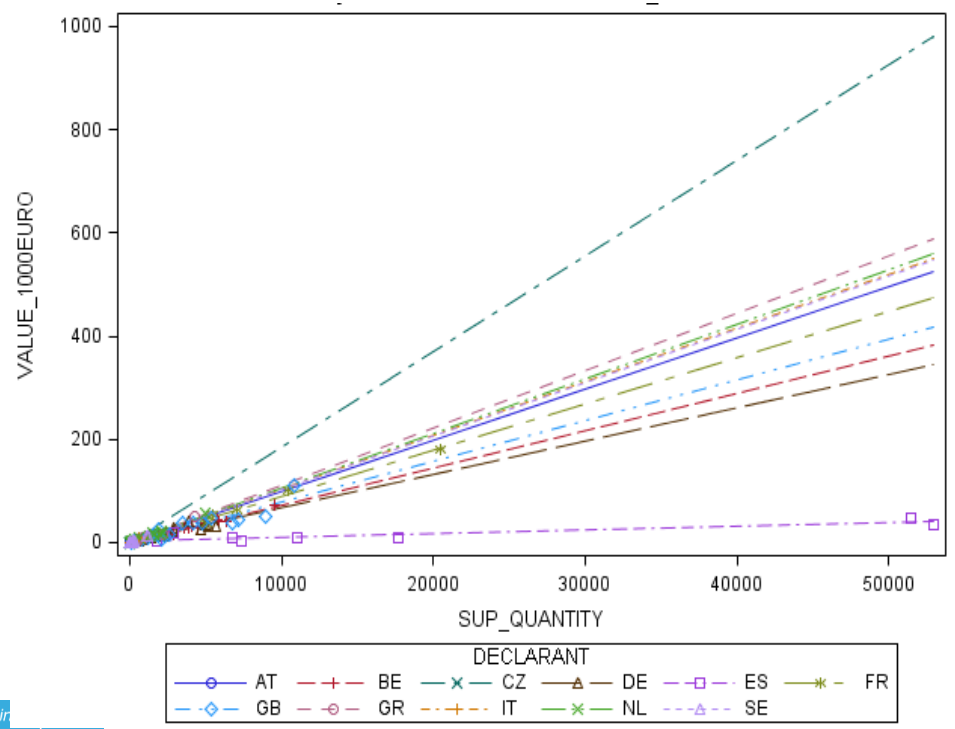
Product	Flow Type	Origin	Destination	Estimated fair price	Estimated fair price interval	Number of observations	Goodness of fit	Outliers detected
03062210	IMP	CA	ES	7.17	(6.82 ; 7.52)	36	0.98	
03062210	IMP	CA	FR	9.67	(9.37 ; 9.97)	35	0.99	1
03062210	IMP	CA	NL	10.51	(10.02 ; 11.00)	35	0.98	1
03062210	IMP	CA	IE	10.77	(9.99 ; 11.56)	16	0.98	2
03062210	IMP	CA	IT	10.90	(10.30 ; 11.50)	36	0.97	
03062210	IMP	CA	BE	11.07	(10.47 ; 11.66)	36	0.98	
03062210	IMP	CA	GB	11.17	(10.43 ; 11.92)	36	0.96	
03062210	IMP	CA	DE	11.20	(10.56 ; 11.84)	36	0.97	
03062210	IMP	CA	PT	13.04	(12.31 ; 13.76)	27	0.98	
03062210	IMP	CA	SE	13.28	(12.76 ; 13.80)	34	0.99	2
03062210	IMP	CA	RO	13.42	(12.21 ; 14.62)	9	0.99	
03062210	IMP	CA	FI	13.98	(11.61 ; 16.34)	11	0.95	
03062210	IMP	CA	DK	14.09	(13.47 ; 14.71)	36	0.98	
03062210	IMP	CA	PL	14.82	(14.10 ; 15.53)	29	0.98	
03062210	IMP	CA	MT	15.38	(14.23 ; 16.52)	33	0.96	
03062210	IMP	CA	GR	15.69	(15.11 ; 16.26)	36	0.99	
03062210	IMP	CA	AT	15.75	(15.20 ; 16.31)	30	0.99	2
03062210	IMP	CA	CY	16.52	(15.14 ; 17.90)	27	0.96	
03062210	IMP	CA	BG	16.58	(4.57 ; 28.60)	3	0.95	
03062210	IMP	CA	CZ	16.62	(15.88 ; 17.37)	36	0.98	

Export to Excel | Page 1 of 1

Testing the heterogeneity of slopes to each PO

For a given Product (P) and country of Origin (O):

2. The heterogeneity of the fitted slopes (the fair prices) is tested



Testing the heterogeneity of slopes to each PO

For a given Product (P) and country of Origin (O):

2. The heterogeneity of the fitted slopes (the fair prices) is tested

More formally, the model is:

$$V_i = \beta_j Q_i + \varepsilon_i$$

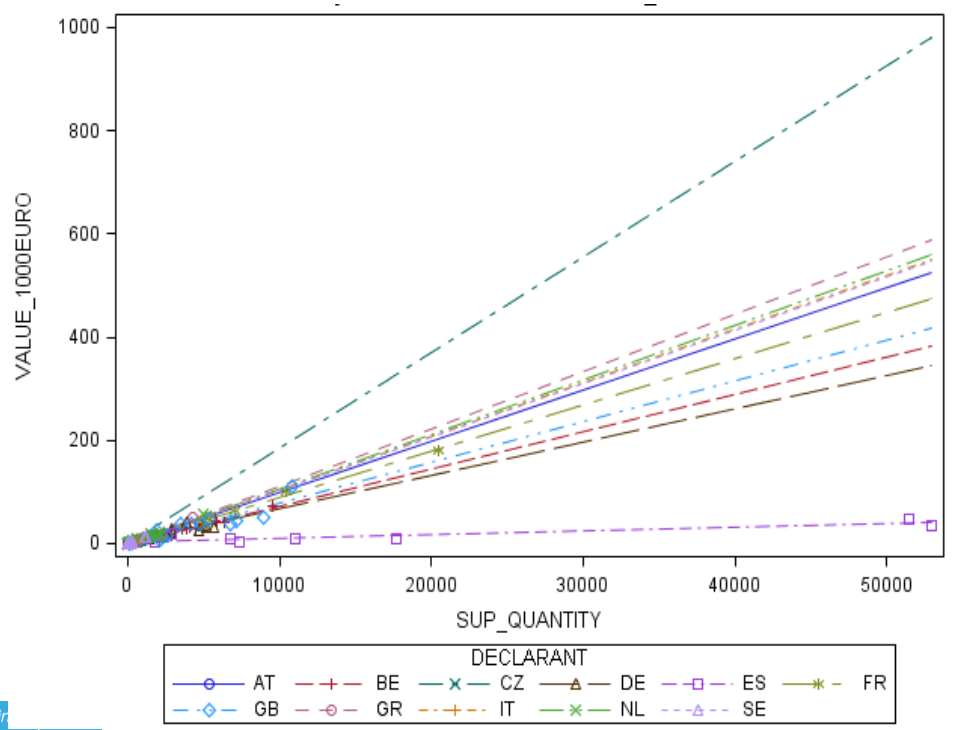
$$i = 1, 2, \ldots, n_j$$

$$j = 1, 2, \ldots, k$$

for k Member States and being n_j the number of flows for MS j

Then, we test that

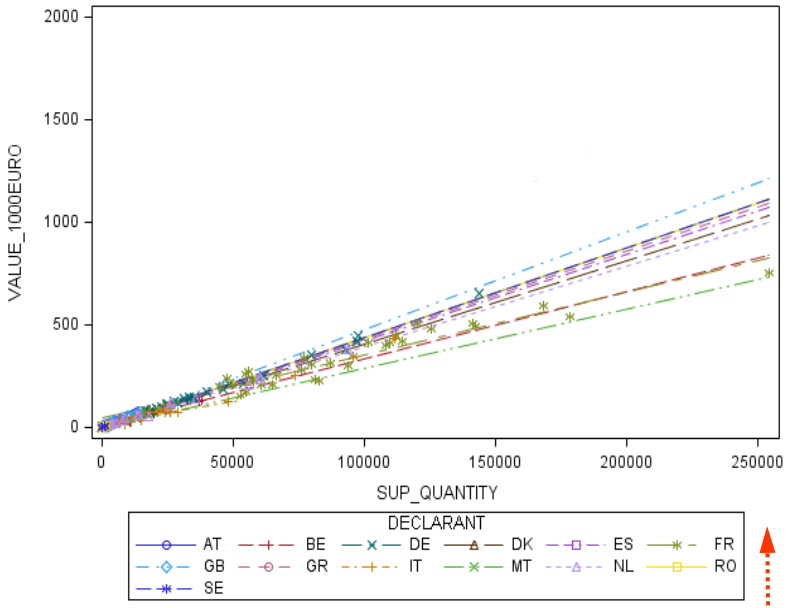
$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$$



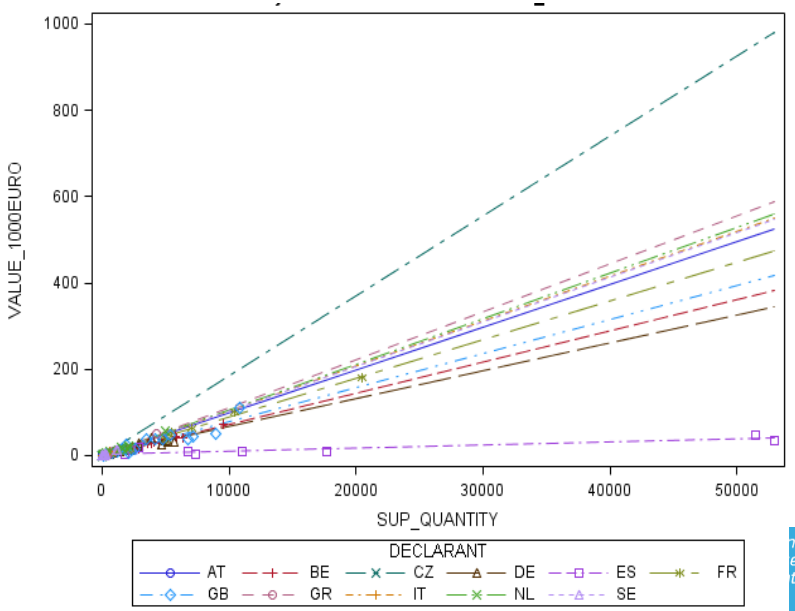
Testing the heterogeneity of slopes to each PO

For a given Product (P) and country of Origin (O):

3. If the hypothesis is not rejected and, thus, the slopes are considered homogeneous, the PO combination is disregarded



this PO is disregarded:
there is an "EU price"

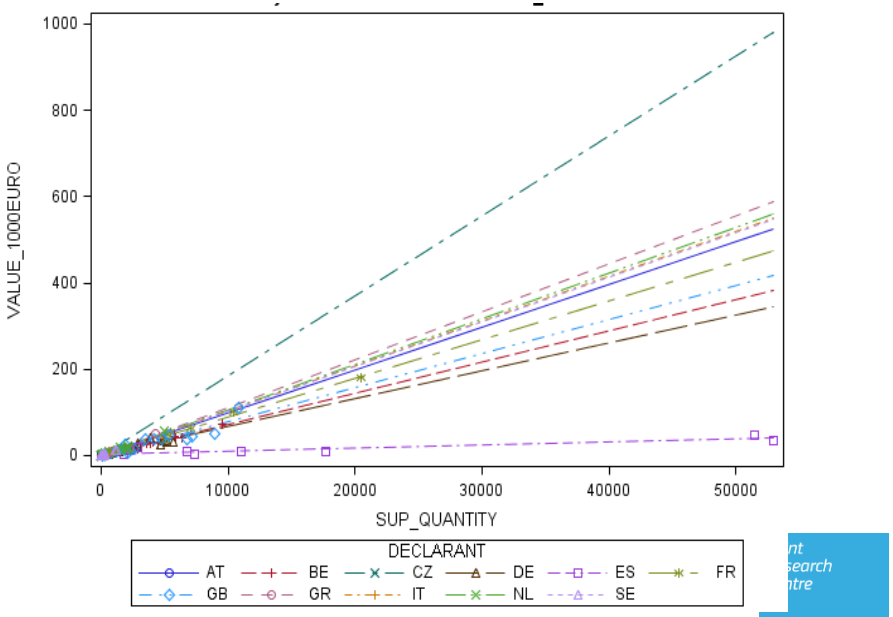


this PO is kept for further analysis
there might be a systematic under pricing

Testing the heterogeneity of slopes to each PO

For a given Product (P) and country of Origin (O):

3. If we conclude that the slopes are heterogeneous, what to do next?



Use of the confidence intervals on the fair prices

Estimated Fair Prices - Thirteenth COMEXT download (01/07/2007 - 30/06/2010), extracted in October 2010

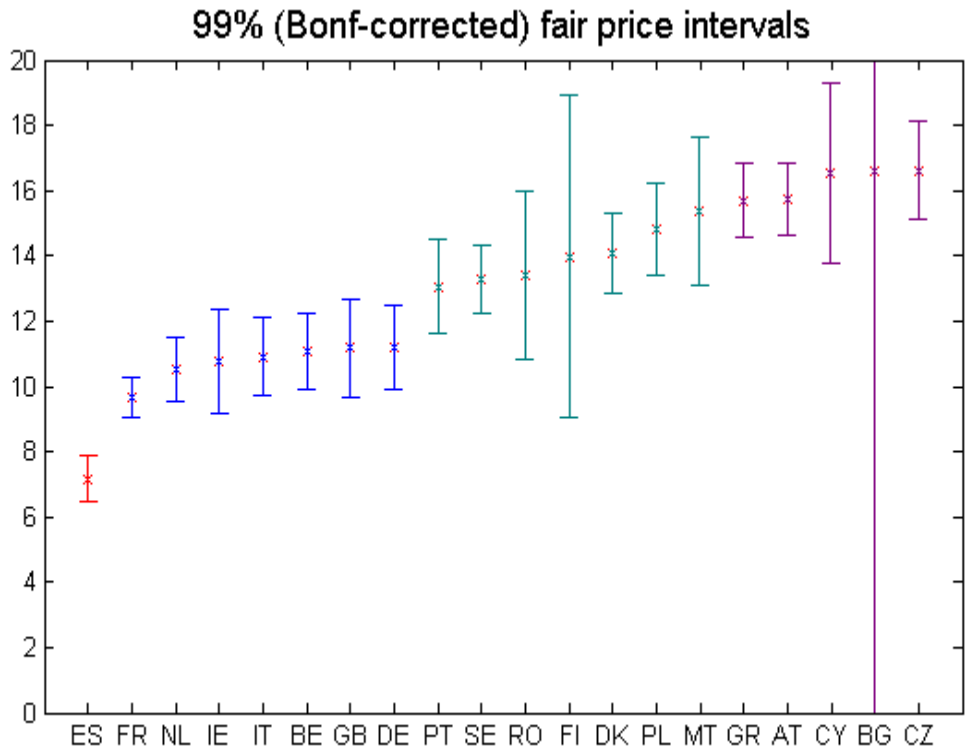
Clear all Filters | Filtered for: **'Product'** starts with **'03062210'** and **'Origin'** starts with **'CA'**

Product	Flow Type	Origin	Destination	Estimated fair price	Estimated fair price interval	Number of observations	Goodness of fit	Outliers detected
03062210	IMP	CA	ES	7.17	(6.82 ; 7.52)	36	0.98	
03062210	IMP	CA	FR	9.67	(9.37 ; 9.97)	35	0.99	1
03062210	IMP	CA	NL	10.51	(10.02 ; 11.00)	35	0.98	1
03062210	IMP	CA	IE	10.77	(9.99 ; 11.56)	16	0.98	2
03062210	IMP	CA	IT	10.90	(10.30 ; 11.50)	36	0.97	
03062210	IMP	CA	BE	11.07	(10.47 ; 11.66)	36	0.98	
03062210	IMP	CA	GB	11.17	(10.43 ; 11.92)	36	0.96	
03062210	IMP	CA	DE	11.20	(10.56 ; 11.84)	36	0.97	
03062210	IMP	CA	PT	13.04	(12.31 ; 13.76)	27	0.98	
03062210	IMP	CA	SE	13.28	(12.76 ; 13.80)	34	0.99	2
03062210	IMP	CA	RO	13.42	(12.21 ; 14.62)	9	0.99	
03062210	IMP	CA	FI	13.98	(11.61 ; 16.34)	11	0.95	
03062210	IMP	CA	DK	14.09	(13.47 ; 14.71)	36	0.98	
03062210	IMP	CA	PL	14.82	(14.10 ; 15.53)	29	0.98	
03062210	IMP	CA	MT	15.38	(14.23 ; 16.52)	33	0.96	
03062210	IMP	CA	GR	15.69	(15.11 ; 16.26)	36	0.99	
03062210	IMP	CA	AT	15.75	(15.20 ; 16.31)	30	0.99	2
03062210	IMP	CA	CY	16.52	(15.14 ; 17.90)	27	0.96	
03062210	IMP	CA	BG	16.58	(4.57 ; 28.60)	3	0.95	
03062210	IMP	CA	CZ	16.62	(15.88 ; 17.37)	36	0.98	

Export to Excel | Page 1 of 1

Use of the confidence intervals on the fair prices

MS are grouped together if their fair price confidence intervals overlap, so that in a cluster the maximum lower bound is smaller than the minimum upper bound



European Commission

The concept of cluster price

P	O	D	cluster countries	POD fair price	cluster price	cluster R2
3062210	CA	ES	ES	7.17	7.17	0.97976
3062210	CA	FR	DE_GB_BE_IT_IE_NL_FR	9.67	10.82	0.97503
3062210	CA	NL	DE_GB_BE_IT_IE_NL_FR	10.51	10.82	0.97503
3062210	CA	IE	DE_GB_BE_IT_IE_NL_FR	10.77	10.82	0.97503
3062210	CA	IT	DE_GB_BE_IT_IE_NL_FR	10.9	10.82	0.97503
3062210	CA	BE	DE_GB_BE_IT_IE_NL_FR	11.07	10.82	0.97503
3062210	CA	GB	DE_GB_BE_IT_IE_NL_FR	11.17	10.82	0.97503
3062210	CA	DE	DE_GB_BE_IT_IE_NL_FR	11.2	10.82	0.97503
3062210	CA	PT	MT_PL_DK_FI_RO_SE_PT	13.04	13.36	0.98693
3062210	CA	SE	MT_PL_DK_FI_RO_SE_PT	13.28	13.36	0.98693
3062210	CA	RO	MT_PL_DK_FI_RO_SE_PT	13.42	13.36	0.98693
3062210	CA	FI	MT_PL_DK_FI_RO_SE_PT	13.98	13.36	0.98693
3062210	CA	DK	MT_PL_DK_FI_RO_SE_PT	14.09	13.36	0.98693
3062210	CA	PL	MT_PL_DK_FI_RO_SE_PT	14.82	13.36	0.98693
3062210	CA	MT	MT_PL_DK_FI_RO_SE_PT	15.38	13.36	0.98693
3062210	CA	GR	CZ_BG_CY_AT_GR	15.69	15.75	0.98819
3062210	CA	AT	CZ_BG_CY_AT_GR	15.75	15.75	0.98819
3062210	CA	CY	CZ_BG_CY_AT_GR	16.52	15.75	0.98819
3062210	CA	BG	CZ_BG_CY_AT_GR	16.58	15.75	0.98819
3062210	CA	CZ	CZ_BG_CY_AT_GR	16.62	15.75	0.98819

Results from the 13th download (July 2007 – June 2010)

24947 combinations of Product and Origin in the download

1513 combinations of Product and Origin with at least 10 importing MS of which:

- 529 turned out to be homogeneous in terms of slopes: for each of these PO combinations it makes sense to define a unique fair EU-price
- 984 are not homogeneous in terms of slopes: for each of these PO combinations we can group together MSs satisfying the criterion of the overlapping confidence intervals, and for each of these groups define a unique fair group-price.

European Commission

Results from the 13th download (July 2007 – June 2010)

Top ten ranked cases taken from the table of the low-priced groups with:

1. focus on the clusters with only one country (filter for *n_country_clusters* = 1)
2. focus on the operationally more interesting cases (filter for *signal score* = 4)
3. filtered cases sorted by descending volume (*Total quantity, for the POD and overall period*)

Sorted by *'Signal score' descending* and by *'Total quantities, for the POD and overall period'* descending

Filtered for: *'n_cluster_countries'* equal 1

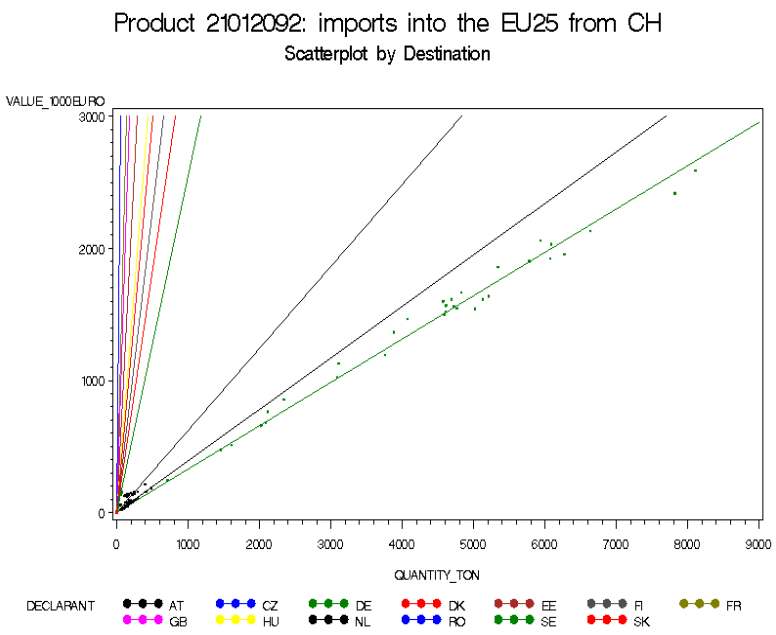
P	O	D	Total values, for the POD and overall period	Relative percentage of the total values	Total quantities, for the POD and overall period	Relative percentage of the total quantities	Total values, for the PO and overall period	Total quantities, for the PO and overall period	Signal score	Fair Price	cluster_countries	n_cluster_countries	n_clusters	cluster_price	cluster_R2	Scatterplot	Cluster Scatterplot	Box plot	Export Helpdesk	TARIC consult
21012092	CH	DE	47259.28	87.90%	141297.9	93.41%	53763.1	151263.4	4	0.3283	DE	1	7	0.3283	0.9966				>	>
08054000	TR	RO	16356.93	16.90%	67544.3	31.62%	96785.21	213592.1	4	0.2444	RO	1	4	0.2444	0.9868				>	>
07031019	TR	BG	3013.85	26.98%	46772	55.21%	11170.86	84720.3	4	0.0612	BG	1	6	0.0612	0.9996				>	>
07049010	MK	RO	1435.94	11.67%	23354.3	30.02%	12305.98	77806.2	4	0.058	RO	1	5	0.058	0.9955				>	>
07133390	EG	RO	4272.87	30.03%	21407.5	52.07%	14226.34	41114.4	4	0.1981	RO	1	4	0.1981	0.9644				>	>
16041319	TH	RO	9259.39	36.42%	9449.3	48.14%	25425.12	19629	4	0.9805	RO	1	3	0.9805	0.9782				>	>
07099090	TR	RO	1006.39	11.45%	9226.7	51.01%	8789.93	18088	4	0.1102	RO	1	8	0.1102	0.9993				>	>
05119190	US	MT	2864.18	14.65%	6486.3	87.44%	19548.66	7418.4	4	0.4183	MT	1	5	0.4183	0.992				>	>
08071100	TR	RO	667.68	5.04%	6485.8	13.54%	13249.48	47887.6	4	0.0993	RO	1	4	0.0993	0.9998				>	>
03061350	VN	IT	17645.02	5.90%	5564.2	9.83%	299190.15	56589	4	3.1239	IT	1	4	3.1239	0.9826				>	>

1 - 10 of 145 Rows per page: 10

Results from the 13th download (July 2007 – June 2010)

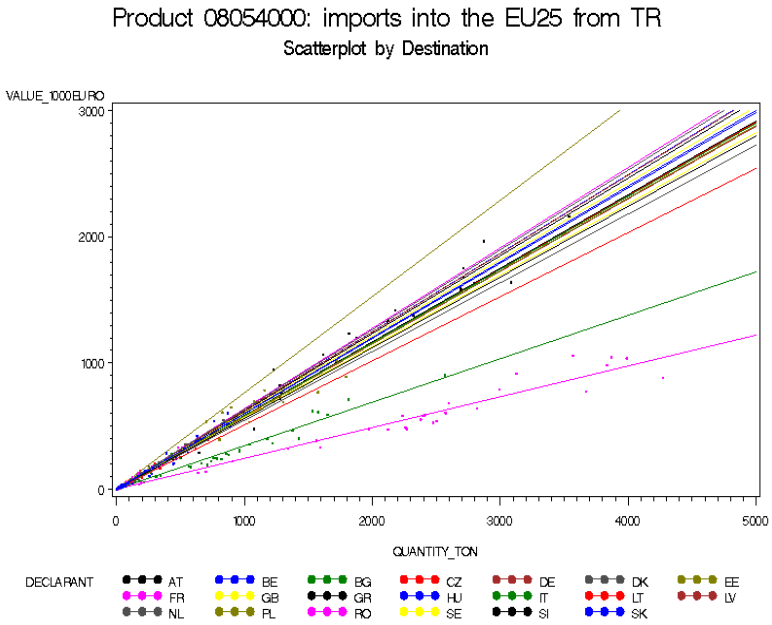
First ranked: 21012092 from CH to DE

Prepared foodstuff, edible preparations



Second ranked: 08054000 from TR to RO and BG

Edible fruit and nuts, grapefruit including pomelos, ...



Joint Research Centre

Summary, caution statements, further work

- Detect relatively few patterns in numerous and diverse fraud control problems.
- Statistical methods used:
  1. To detect spikes.
  2. To detect price outliers.
  3. To prioritize outliers (in two ways).
  4. To detect systematic under-declarations.
- Applicability to both aggregated and dis-aggregated trade data.
- Estimation of fair prices; an important spin-off of the detection of price outliers procedure

Summary, caution statements, further work

- Different methods to detect outliers and spikes:
  1. Iterative backward search (BS)
  2. Forward Search (FS)
  3. Traditional robust methods (LTS, LMS, …)
- Importance of controlling number of false positives taking into account multiplicity of comparisons.
- Presence of mixtures of prices for the same product, origin and destination