PLISKA
STUDIA MATHEMATICA BULGARICA

ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **18** (2007), 271–292

PLISKA
STUDIA MATHEMATICA
BULGARICA

EXTRACTION OF FRAUD SCHEMES FROM TRADE SERIES

Charalambos Moussas   Veska Noncheva

It is very often the case that the patterns of a fraudulent activity in trade are hidden within existing trade data time series. Furthermore, with the advent of powerful and affordable computing hardware, relatively big amounts of available trade data can be quickly analyzed with a view to assisting anti-fraud investigations in this field. In this paper, based on the availability of such import/export data series, we present a statistical method for the identification of potential fraud schemes, by extracting and highlighting those cases which lend themselves to further investigation by anti-fraud domain experts. The proposed method consists in applying time series analysis for prediction purposes, calculating the resulting significant deviations, and finally clustering time series with similar patterns together, thus identifying suspect or abnormal cases.

1. Introduction

There is an increasing interest in the application of statistical data analysis methods in the area of anti-fraud in general. This is mainly due to the wide availability of big amounts of related data in electronic form, and the need to turn this data into useful information in order to be able to detect fraudulent activities as soon as possible.

Information on the trade of goods is typically made available through both national and international statistical offices. In this respect, their role is to collect,

2000 *Mathematics Subject Classification:* 62H30, 62M10, 62M20, 62P20, 94A13
Key words: Fraud Detection, Time Series Analysis, Forecasting, Cluster Analysis

process, and disseminate import/export indicators, such as the quantity and the value of the trading goods. One of the main sources of this kind of information is the national customs authorities. Thus, data is usually collected during customs procedures and subsequently forwarded to the appropriate statistical offices, on a monthly basis. As a result, monthly import/export data can be made available in the form of monthly trade series which can be further analyzed, by using time series analysis, with a view to discovering trends matching suspected fraud patterns.

The rationale behind our approach is that although significant changes in the amount of trading goods can be due to the market evolution, there is always a number of cases where this behavior could be a sign of fraud. Our objective is therefore to detect potential fraud schemes by identifying fraud-like, or abnormal, patterns in the underlying data. Whether any extracted schemes can be actually associated to suspected or established fraud will be based on further investigation by anti-fraud domain experts.

Statistical fraud detection approaches may be 'supervised' or 'unsupervised'. In supervised approaches, samples of both fraudulent and non-fraudulent records are used to construct models which allow one to assign new observations into one of the two classes. Of course, this requires one to be confident about the true classes of the original data used to build the models. It also requires that one has examples of both classes. Furthermore, it can only be used to detect frauds of a type which have previously occurred. In contrast, unsupervised approaches do not require training samples with fraudulent and non-fraudulent cases and can be effective even for new types of fraud. The unsupervised approaches usually seek those cases which are most dissimilar from the norm. These can then be examined more closely. Outliers are a basic form of non-standard observation.

A major assumption in our unsupervised modeling approach is that the future behaves as the past, which implies that it can be described by the same mathematical model. We also separate the available trade data series into a *historical data* part and a *present data* part. Then, based on the past trade history, a model of normal behavior is derived which is then compared to the present trade data. As a result, significant changes which are not consistent with the model are identified and further analyzed.

Our method represents an improvement of the method discussed in [10], where an application to fraud detection in external trade has been considered. Other data analysis tools have been applied successfully to detect activities such as money laundering, e-commerce credit card fraud, telecommunications fraud, computer intrusion and medical insurance. Some statistical tools, and the areas in which statistical fraud detection technologies are most used, are discussed in [4]. Artis, Ayuso and Guillen have described an approach to modeling fraud behavior in car insurance ([1], [2]). Fanning, Cogger and Srivastava ([8]), and Green, Calderon and Choi ([6], [7]) have examined some classification methods for detecting management fraud. Fraud detection tools have also been applied to sporting events ([3], [13]).

The rest of the paper is organized as follows: In section 2, we introduce the problem addressed in this paper. In section 3 the proposed unsupervised statistical approach to fraud detection is described, while in section 4 some important implementation issues are considered. Section 5 gives an example and section 6 concludes the paper.

2. Problem Formulation

One of the most frequent types of fraud encountered in Customs is the false declaration of origin concerning the import of a product into the European Union (EU) market ([12]). There exist more than one reasons for declaring a false origin for the trading goods, such as the circumvention of anti-dumping duties, or the use of a preferential trade regime of another EU partner, or the existence of a specific EU import quota which has been reached, etc. In all these cases, the fraud consists in declaring that the product under consideration comes from an EU partner other than the one actually exporting it. Thus, the fraud mechanism typically includes an initial export from the first non-EU country, say country A, to a second non-EU country, say country B, and the subsequent export from the second non-EU country B towards one of the EU member states (see Figure 1).

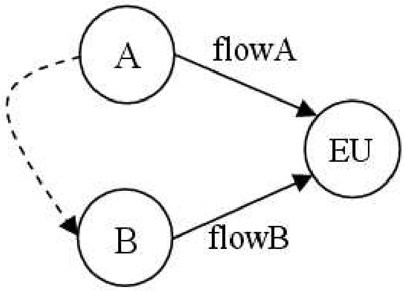

Figure 1: Simple fraud mechanism

To be more specific, let us assume that the fraud consists in the circumvention of anti-dumping duties which have been put in place at a certain point in time, and for a specific EU partner country. That is, after a certain date, the duties paid when importing a specific product from this partner into the EU market are considerably higher than what it used to be. As a result, and in order to avoid the extra money being paid, thus making the product more competitive on the market, the EU partner could first export it to another non-EU country for which no extra duties are imposed on the product under consideration, and then from that country to the EU. In fact, very often more than one such "transfers" between EU partners may take place in order to further hide the actual flow of products. Thus, the trading goods might go through more than one non-EU countries before actually reaching the EU borders, either in a sequential order or in parallel, as illustrated in Figures 2 and 3 respectively, or even a combination of the two.

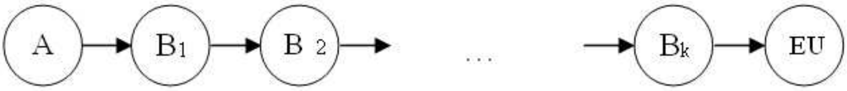

Figure 2: Sequential fraud mechanism

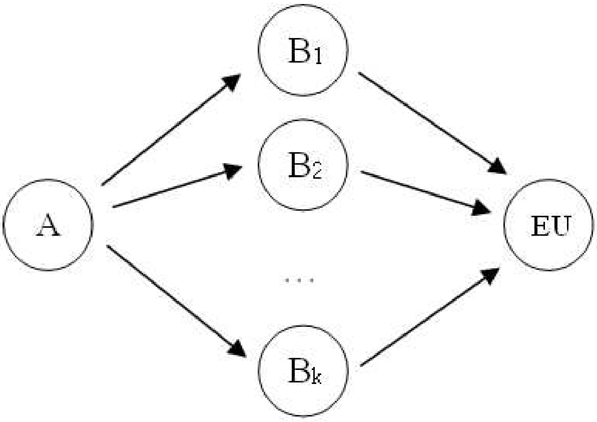

Figure 3: Parallel fraud mechanism

Our goal is trying to detect this sort of fraudulent situations based on the trading quantities. The information at our disposal is the trade information concerning the various trade flows between the different countries involved. As far as the EU trade is concerned, specific imported and exported quantities are available for all EU partners on a monthly basis. This implies that for the case of Figure 1 both $flowA$ and $flowB$ related time series will be available for analysis, while in the case of Figure 2 and 3, some intermediate trade flows, between non-EU countries might not be known. In any case, only the known time series are taken into account. The period of the available data under consideration will include of course the specific point in time in which the anti-dumping duties are introduced. Now assume for a moment that the situation is as depicted in Figure 1. If the above mentioned fraud mechanism applies, then the monthly exports from country A into the EU, that is $flowA$, should drop shortly after, say with a delay $\mathbf{d}$, the import duties have been introduced. On the other hand, the monthly exports from country B into the EU market, that is $flowB$, should rise as a result of the "transfer" from country A to country B. This rise in EU imports from country B takes place say after a time delay $\mathbf{D}$. This situation is depicted in Figure 4, with $\mathbf{d}$ and $\mathbf{D}$ being the two time delays where, typically, $\mathbf{d}$ is less than $\mathbf{D}$. In case, of course, that more countries are involved as shown in Figures 2 and 3, then more such increases could be actually happen, both between non-EU countries in the case of Figure 2, and between EU and its partners in the case of Figure 3. The corresponding time delays will vary, depending on the individual case. Therefore, the fraud detection problem we are dealing with can be summarized as follows: based on the available data flows between different countries, in the form of monthly trade series, for a specific product, our objective is to identify groups of trade flows following fraudulent patterns as explained above and is illustrated in Figure 4 for the simplest case.

It is worth mentioning here that although reliable trade data for the EU member states are readily available through the statistical office of the European Union, or EUROSTAT, this is definitely not the case for trade information between non-EU countries. As a result, in the examples that follow, the analysis will be based on the declarations of the EU member states only, and therefore no trade information between non-EU countries will be taken into account. Note, however, that the proposed method applies equally well to any kind of available time series. It is just a matter of dealing with as much reliable information as possible. For example, in the simple case of Figure 1, it is obvious that if the trade flow from country A to country B were also available, it should then typically follow a pattern similar to that of $flowB$. Thus, in this simple case, being

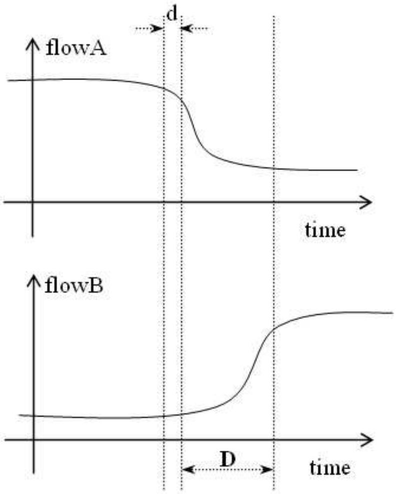

Figure 4: Opposite trade patterns

able to identify all three flows within the same group thus closing the “loop”, would be a strong indication of suspected fraud under the above mentioned fraud scheme.

3. Method

The fraud scheme extraction method we propose, eventually generates sets of warning signs which when appropriately grouped together may lead into likely fraud schemes. As we have already mentioned, the actual verification of committed fraud needs further anti-fraud investigations.

We follow an unsupervised statistical approach. We assume that the available trade series represent realizations of random processes. Thus, for each trade series there is a specific probability model describing the observations, and any fraudulent behavior will typically cause the corresponding model to change. Based on the observations we use time series analysis, as well as clustering, in order to model the following concepts: Trade Flows, Warning Signs, and Fraud Schemes.

3.1. Trade Flow Modeling

We consider the available import export data between the various countries as a collection of data flows that need to be examined. Time series data often has a pattern which repeats every s-time periods. Monthly data may have seasonal period $s = 12$. The import/export of a product from one country to another may repeats itself every 12 months. Randomness in the seasonal patterns from one cycle to the next is possible. In other words, the import/export quantities are not so much related from month to month as they are from year to year. The question is how to model the import/export series so that dependence both on neighboring months and on months from the previous years to be taken into account.

In order to model the stochastic mechanism that gives rise to the observed data over time we use time series analysis based on sARIMA (seasonal Auto Regressive Integrated Moving Average) models, known as the Box-Jenkins approach. Box-Jenkins methodology is a class of linear time series forecasting techniques that capture the linear dependency of the future values on the past values. They are able to model a wide spectrum of time series behavior. A sARIMA$(p, d, q)(P, D, Q)_s$ model includes the following types of parameters: the autoregressive parameters (p, P), the number of differencing passes (d, D), the moving average parameters (q, Q), and the seasonal period s.

The $\{X_t\}$ process is an sARIMA$(p, d, q)(P, D, Q)_s$ if the differenced series $Y_t = (1 - B)^d (1 - B^s)^D X_t$ is a process defined by $\varphi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t$, $\{Z_t\} \in WhiteNoise(0, \sigma^2)$, where d and D are nonnegative integers, $\varphi(z) = 1 - \varphi_1 z - \cdots - \varphi_p z^p$, $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_p z^p$, $\theta(z) = 1 - \theta_1 z - \cdots - \theta_q z^q$, $\Theta(z) = 1 - \Theta_1 z - \cdots - \Theta_q z^q$, and both $\varphi(z) \neq 0$ and $\Phi(z) \neq 0$ for $[z] \leq 1$ (see [5]).

In modeling real data it might not be reasonable to assume that the seasonal component repeats itself precisely in the same way cycle after cycle. Seasonal ARIMA models allow for randomness in the seasonal pattern from one cycle to the next.

In a traditional ARIMA methodology the user must specify the model. The determination of an appropriate sARIMA$(p, d, q)(P, D, Q)_{12}$ model to represent an observed time series involves a number of interrelated problems. A distinctive feature of the data that suggests the appropriateness of an ARIMA model is the slowly decaying positive sample autocorrelation function. Trend and seasonality are also characterized by autocorrelation functions that are slowly decaying and nearly periodic, respectively.

It is not advantageous from a forecasting point of view to choose the autoregressive and the moving average parameters arbitrarily large. Fitting a very high order model will generally result in a small estimated white noise variance, but when the fitted model is used for forecasting, the mean squared error of the forecasts will depend not only on the white noise variance of the fitted model but also on errors arising from estimation of the parameters of the model. These will be larger for higher-order models ([5]).

Thus the traditional model selection process typically requires expert experience. In our approach the model is automatically selected. Our prime criterion for parameters selection is the Akike Information Criterion (AIC), defined as $AIC = -2lnL_x + 2n_{par}$, where lnL_x is the log-likelihood value and n_{par} represents the number of parameters in the fitted model (see [15]). We choose the fitted model with smallest AIC value. A small difference in AIC value (less than 2) between two satisfactory models may be ignored in the interest of model simplicity.

The estimates of the noise in a probability model are the residuals. If there is no dependence between the residuals we estimate their mean and variance. If there is significant dependence among the residuals we look for a more complex stationary time series model. Dependence means that past observations can assist in predicting future values. Final selection of the model depends on the results from a variety of goodness of fit tests, such as Ljung-Box test for independence, Shapiro-Wilk test for normality and the sample autocorrelation function of the residuals. If the residuals are not consistent with their expected behavior under the minimum AIC model then competing models should be checked until we find a model that passes the goodness of fit tests.

Trade data flows can be further analyzed with a view to identifying warning signs as we explain below.

3.2. Warning Sign Modeling

A fraud detection method for trade, should somehow be able to assign either a suspect or a normal profile to trade time series. Forecasting techniques help us to understand past events, discern patterns and project those patterns into the future. Thus, we can perform time series forecasting and identify warning signs, pointing to a suspect behavior.

Let n be the total number of observations at a given time. We select the first k observations $X_1, X_2, \ldots, X_k, k < n$, from the data series, and for each import trade flow we derive a model of normal behavior based on this trade history represented by the selected observations. Then we compute the forecast values $X_i^{forecast}, i = k+1, \ldots, n$, by applying the model derived from the historical data. We also know the observed values $X_i, i = k+1, \ldots, n$, and we shall define the notion of warning sign in such a way as to indicate whether the predicted values are significantly different from the corresponding observed ones. Significant difference means that the observed value X_i is outside the $\gamma\%$ prediction interval, denoted as (a^i_γ, b^i_γ). Usually the level of confidence is $\gamma = 95$. Thus, we define the warning sign WS_i for each observed value X_i, $i = k+1, \ldots, n$ in the following way:

$$WS_i = \begin{cases} X_i - b^i_\gamma & \text{if } X_i > b^i_\gamma \\ X_i - a^i_\gamma & \text{if } X_i < a^i_\gamma \\ 0 & \text{if } a^i_\gamma \leq X_i \leq b^i_\gamma \end{cases}$$

where (a^i_γ, b^i_γ) is the γ% prediction interval.

As illustrated in Figure 5, for the simplest type of fraud scheme considered in Figure 1, it is expected that both *flowA* and *flowB* will lead to non-zero warning signs whose opposite sign indicate the possibility of an underlying fraudulent situation, as explained in section 2.

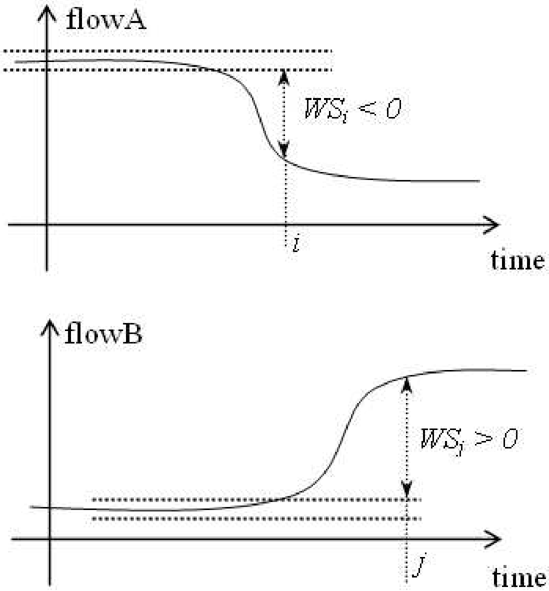

Figure 5: Warning signs

In fact, once such warning signs are detected, they can be further analyzed with a view to identifying potential fraud schemes. Our approach on how to proceed further is explained in the next subsection.

3.3. Fraud Scheme Modeling

Recall from section 2, that our goal is to identify groups of trade flows having fraudulent patterns that match the fraud schemes under consideration. Since the warning signs we defined above actually point to suspect trade flows, then it is the combination of different warning signs that can lead to the identification of suspect fraud schemes. Broadly speaking, this combination can be done based either on specific trade fraud scenarios, or on abnormal trade behavior in general. We proceed as follows.

For the observed monthly period $[k+1, n]$ we summarize the behavior of each trade flow, that corresponds to a pattern for the data series, by means of the following characteristics:

- the mean of the historical values $Hm = \frac{1}{k}\sum_{i=1}^{k} X_i$
- the mean of the warning signs $Sm = \frac{1}{n-k}\sum_{i=k+1}^{n} WS_i$, and
- the ratio $R = \begin{cases} \frac{Sm}{Hm}100\% & \text{if } Hm > 1 \\ Sm100\% & \text{if } 0 \leq Hm \leq 1 \end{cases}$

Based on these three variables, we cluster the data series, thus extracting different groups of trade flows with similar patterns, using cluster analysis. The dataset for clustering has the structure $m \times p$ objects-by-attributes matrix, where rows stand for objects (the time series) and columns stand for the three interval scaled variables (Hm,Sm,R) measured on the time series. We apply the Partitioning Around Medoids (PAM) method described in [9] with manhattan metric. It is robust with respect to outliers and other violations of the assumption of spherically normal clusters because it is based on sums of non-squared distances. The overall quality of the partition is measured by the average silhouette width [14]. This index allows a reasonable data structure to be automatically found.

The overall objective during this clustering process is to come up with an appropriate number of clusters which can be further examined, so that either suspect and/or abnormal trade flow combinations can be identified and extracted.

4. Implementation issues

There exist a number of issues, from an anti-fraud point of view, which will greatly affect the pattern of any extracted fraud schemes, and for which expert advice from anti-fraud practitioners, can lead to far better results than making some generic modeling assumptions based on the modeling parameters we just introduced in section 3. A number of these issues are discussed in the sequel.

4.1. Kinds of clusters

Since we are interested in significantly changing patterns, we expect to identify a main large cluster of flows corresponding to normal patterns with zero warning signs, as well as a number of relatively smaller clusters having either negative or positive warning sign values. Thus, we would generally expect five kinds of clusters, which we shall describe as follows:

- *Normal* - Clusters of trade flows having stable patterns
- *Highly Increasing* - Clusters of trade flows having significantly increasing patterns
- *Increasing* - Clusters of trade flows having relatively increasing patterns
- *Highly Decreasing* - Clusters of trade flows having significantly decreasing patterns
- *Decreasing* - Clusters of trade flows having relatively decreasing patterns

Then, for the simple case shown in Figures 1 and 4, $flowA$ would ideally belong to a *Highly Decreasing* cluster, while $flowB$ would rather be in a *Highly Increasing* cluster. Furthermore, the fact that the two flows are related through the same importing party (the EU, in this case), can be used to identify and assign the two flows into a common context of a suspected fraud scheme of this sort.

The same procedure can be applied in a straightforward way to the more general cases illustrated in Figures 2 and 3. Of course, this depends on the availability of the corresponding trade data and, as we already mentioned in section 2, it might be difficult to obtain reliable data for trade between non-EU countries. However, by clustering all available trade flows into a maximum of 5 kinds of clusters, as explained above, and by appropriately combining the results through any common parties involved, more sophisticated fraud schemes can be identified.

4.2. Aggregate information

It is often the case that a specific fraud scheme might concern transfers of goods that can be better identified and explained if one considers a given set of countries as a whole, rather than each member individually. This is, for example, the case of EU in the simple example of Figure 1 and 4 where EU comprises the member states of the European Union. Depending on the fraud scheme, it can include the 10 new member states, in which case we would rather talk about EU25, or it can contain only the 15 member states before the 1^{st} of May 2004, thus being referred to as EU15. Of course, if it is likely that the suspected fraud is at an individual country level, then we would include all individual trade flows related to each member state. Note, also, that we can even analyze both individual and aggregate information together, as far as the resulting clusters are correctly interpreted. Therefore, the decision on the specific trade flows to be included in the overall analysis strongly depends on the underlying fraud scheme model, and it should be taken in an early stage since it affects the trade flow modeling as well.

4.3. Prediction period

The period $[k+1, n]$ chosen for the prediction of the trade series plays an important role in the extraction of potential fraud schemes. Typically, we'll assume a default period of six months, $k+1, k+2, k+3, k+4, k+5, k+6$, which both takes care of the problem of wrong declarations between consecutive months due to delays in customs procedures, and represents a fairly wide time frame for detecting significantly changing trade patterns. Thus, the parameter prediction period should be chosen carefully in the beginning of the overall analysis process. The starting point $k+1$ is chosen as the month next to the one which might have triggered such an anti-fraud analysis in the first place. For example, in the anti-dumping case mentioned in section 2, it could represent the month of the introduction of the extra duties.

Another issue closely related to the prediction period, is how one deals with new data becoming available after time n. The approach we take in this paper is the following. We repeat the trade flow modeling phase by including the observed values $X_{k+1}, X_{k+2} \ldots, X_{k+6}$ for the months $k+1, k+2, k+3, k+4, k+5, k+6$ into the historical set of values, and considering the new prediction period $[k+7, k+12]$. New warning sign variables are calculated and clustering analysis is applied again, thus resulting in a new set of clusters subject to interpretation. Following the same approach for each new set of observed values, allows us to identify significant changes appearing later than the initial prediction period, but which might still relate to the possible triggering event back in time k.

We should finally mention here that, as far as the clustering phase is concerned, one can even cluster together the overall results coming from the different prediction periods under consideration. By doing so, one can get clusters of trade flows with similar characteristics but related to different prediction periods, thus allowing one to extract potential fraud patterns spread over a more extended period of time.

5. Example

The proposed method is applied to a real investigation case regarding exports of large quantities of a specific product from a specific non-EU country towards the rest of the world, thus including the EU member states. For confidentiality purposes, they will be referred to as *Product* and *Partner1*, respectively. The main objective is establishing to what extent (if any) the Product is deviated through other non-EU countries before entering the EU territory, thus benefiting by more favorable duty rates applicable there. In fact, a quantitative upper limit (quota) to EU imports coming from *Partner1* has been set up in an annual basis, so that quantities exceeding this limit are subject to duty rates which are approximately ten times higher. Furthermore, because of the enlargement process in the European Union that took place on May 1^{st}, 2004, exports from *Partner1* to the 10 new member states before and after that date could be examined, given the possibility that the same goods can be re-exported from the new member states to the old ones after the accession, thus benefiting from zero duty rates due to the EU common market. In both these cases, the underlying trade activities might involve some fraudulent behavior in the sense of compliance to the EU specific regulations on this matter. Keep in mind, however, that our method is not restricted only to a specific partner country, so that any other EU partner with a significant changing trade flow, leading to highly likely fraud scheme or showing abnormal behavior, will also be identified during the overall extraction process.

We shall now identify the specific trade flows that directly relate to these two slightly different causes of potential fraud, with a view to identifying the underlying data series which will be included in the analysis. First of all, the exports from *Partner1* to the EU15 are of great importance in both cases, and they are available through the Comext database of Eurostat, so that they will be included in the analysis. Note that the EU15 is considered as a whole, since we are interested to the overall EU15 imports and furthermore the member states have the same import duties. The corresponding trade flow, is illustrated in Figure 6.

The exports of all non-EU15 countries towards the EU will also be included, and they are also available in Comext, through the import declarations of the member states. The trade flows from *Partner1* to the rest of non-EU15 countries would be also very useful, and although they could be partially available from other sources, they will not be included in our analysis since their reliability is often questionable. However, because of the enlargement process we mentioned above, 13 candidate countries which will be referred to as CC13 had already started reporting trade information in Comext since 1999. Therefore, also the available trade flows of exports from non-EU15 countries towards these 13 candidate countries (of which 10 represent today the new member states of the EU) will be included in our analysis. Note, that this also includes trade between all the 13 CC's, having different input duty rates before the enlargement date of May 1^{st}, 2004. Figure 7 illustrates all these relevant trade flows which will be taken into account in this example. MS_i, CC_i, and $Partner_i$, denote an EU Member State, a Candidate Country, and a country from the rest of the world, respectively. For the sake of simplicity, each of the arrows we used, contains any existing trade flows for all possible pairs (i, j).

As far as the prediction period is concerned, we shall consider two consecutive 6-month time intervals. The first is the interval from January to June 2004, or [200401, 200406], that is, mainly before the enlargement process and in which the potential fraud could be due to differences in import duty rates between the EU and any other country in general. The historical information will include the observed values in the 2-year period [200201, 200312]. Next, we shall also consider the prediction period [200407, 200412], which is well after the enlargement date. Thus, in this case, potential fraud could be also due to the import duty rates of the old member states, with respect to the new member states, which do not

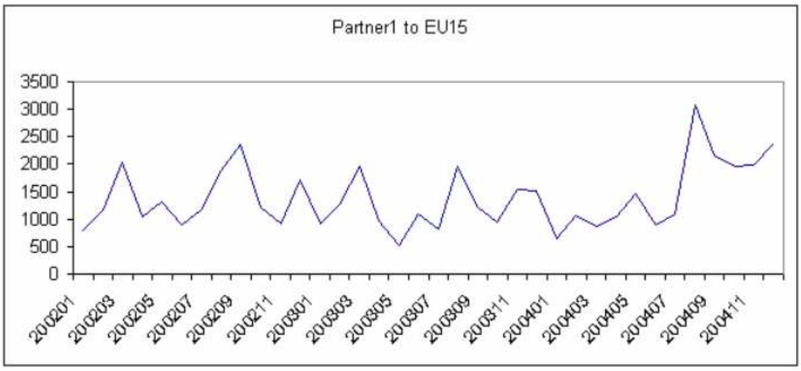

Figure 6: Trade flow from *Partner1* to EU15

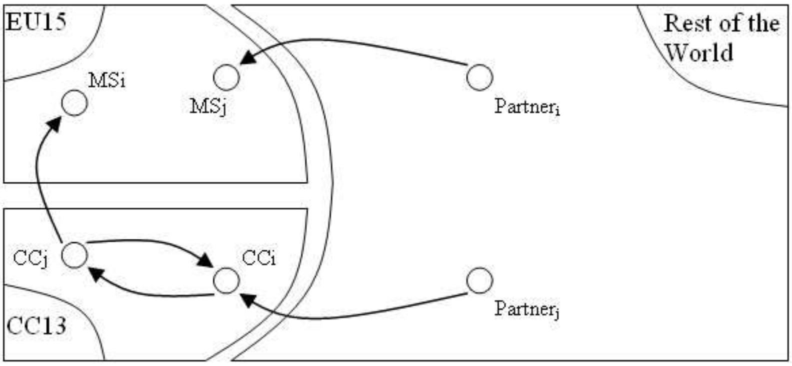

Figure 7: Relevant trade flows

exist anymore. Here, the historical information will include the observed values in the period [200201, 200406] which corresponds to $2 + 1/2$ years.

The implementation is done by means of the **R** statistical software ([11]).

5.1. First Prediction Period

We apply our method for the prediction interval from January to June 2004. The total number of trade flows, in the sense of Figure 7, and for which at least one non-zero monthly quantity appears in the corresponding trade data series during the overall period, from January 2002 to June 2004, is 146. Using the

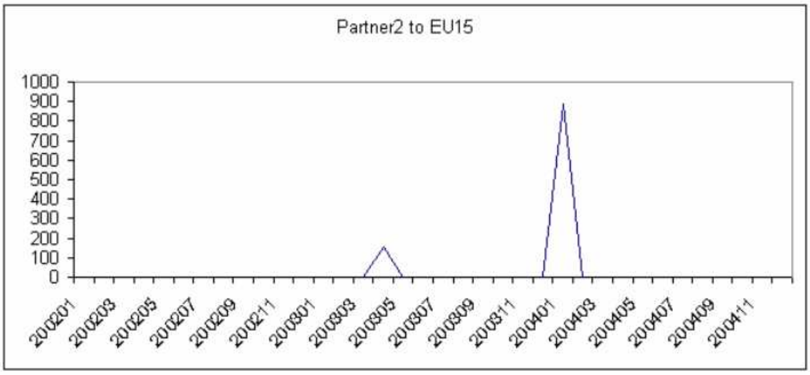

Figure 8: Trade flow from *Partner2* to EU15

PAM method with manhattan metric we find 4 meaningful clusters. A *Normal* cluster of size 132 (i.e. containing 132 data series) with stable patterns, a *Highly Increasing* singleton cluster with the following characteristics

$Exporter$	$Importer$	Hm	Sm	R
$Partner2$	EU15	6.5	117	1800

and its corresponding trade series shown in Figure 8, an *Increasing* cluster of size 8, containing trade flows with the following characteristics

$Exporter$	$Importer$	Hm	Sm	R
******	******	0.004	4.167	416.672
*****	******	1.433	6.375	444.782
*****	******	1.475	6.742	457.074
*****	******	0	10.5	1050
*****	******	0	4.485	448.5
*****	******	0	3.333	333.333
*****	******	0	7	700
*****	******	5.055	13.209	261.331

and, finally, another *Normal* cluster of size 5 with trade flows as follows

$Exporter$	$Importer$	Hm	Sm	R
******	******	1065.638	116.185	10.903
$Partner1$	EU15	1302.642	20.794	1.596
*****	******	1041.138	-134.801	-12.947
*****	******	570.175	-167.818	-29.433
*****	******	804.443	-182.435	-22.678

and which represents the cluster of the "top" exporters as we can see form the variable Hm. (Asterisks, instead of names, have been used for confidentiality purposes only). These "top flows" can have an either positive or negative ratio R, but they are considered as normal because this number is small, compared to that in the previous two clusters. On the other hand, even a small R can imply a big difference in absolute quantities, and this is very important from an anti-fraud point of view, due to the increased value of money involved. Thus, this kind of *Normal* clusters must always be examined thoroughly. Note, in fact, that the second line corresponds to the flow from *Partner1* to EU15, illustrated in Figure 6.

As a result of the clustering process for this first prediction interval, *Partner2* would be pointed out as an emerging exporter whose activity regarding the underlying product must be further examined by antifraud investigators.

5.2. Second Prediction Period

We now apply our method for the prediction interval from July to December 2004. In this case, the total number of available trade flows according to Figure 7, and for which at least one non-zero monthly quantity appears in the corresponding trade data series during the overall period, from January 2002 to December 2004, is 160.

Here, the PAM method with manhattan metric yields 5 clusters. A *Normal* cluster of size 149 with stable patterns, a *Highly Increasing* singleton cluster with the following characteristics

$Exporter$	$Importer$	Hm	Sm	R
$Partner3$	CC1	0	87.5	8750

another *Highly Increasing* singleton cluster with

$Exporter$	$Importer$	Hm	Sm	R
$Partner4$	CC1	0	58.333	5833.333

an *Increasing* cluster of size 5, containing trade flows with the following characteristics

$Exporter$	$Importer$	Hm	Sm	R
******	******	0	31.55	3155
*****	******	0	17.333	1733.333
*****	******	0	17	1700
*****	******	0	13	1300
$Partner5$	CC1	0.02	24.623	2462.297

and, finally, another *Normal* cluster of size 4 with trade flows as follows

$Exporter$	$Importer$	Hm	Sm	R
******	******	1418.923	0	0
$Partner1$	EU15	1241.587	64.119	5.164
$Partner1$	CC1	945.917	-143.846	-15.207
*****	******	843.929	-134.512	-15.939

which, as in the previous case, represents the cluster of the "top" exporters. We notice that, in this case, there is one less trade flow than before, which implies that one flow being close to the border between the two *Normal* clusters, was assigned to the large one.

Consider now the two flows related to Partner1. We see that its exports to EU15 are increasing, while at the same time, those to the candidate country CC1, which is furthermore a new EU member state after May 1^{st} 2004, are decreasing. This combination could suggest a potential fraud scheme as depicted in Figure 1, with country A being Partner1 and country B being CC1. As we mentioned before, although in both cases the ratio R is not very big, the underlying quantities could be of importance because of the large absolute quantities involved. To be more specific, the fraud could consist in exports from *Partner1* towards the EU15 deviated through the new member state CC1, before the prediction period, while the opposite behavior we notice here would imply that the deviation stopped during the prediction period. This opposite behavior is also evident if we compare Figure 9 which illustrates the flow from *Partner1* to CC1, with Figure 6 which shows the flow from *Partner1* to EU15. Finally, in Figure 10, the exports from CC1 to EU15 are also illustrated. Comparing Figure 10 with Figure 9, we see that they both have a peak in August 2003, while exports from CC1 to EU15 have considerably dropped after the accession date. In any case, in order to draw any conclusions, further verification is required by anti-fraud domain experts.

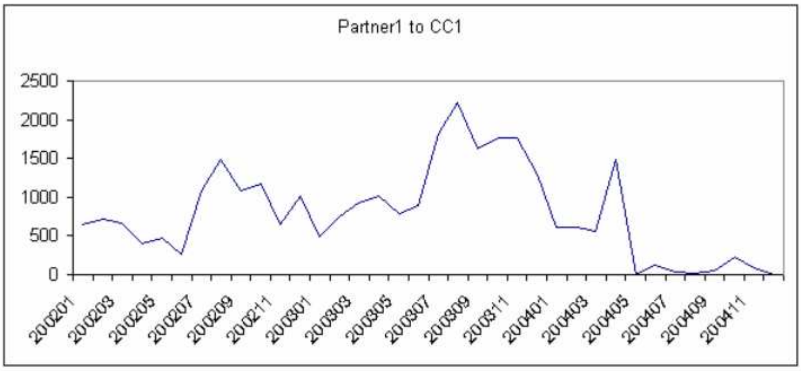

Figure 9: Trade flow from *Partner1* to new member state CC1

Let us now compare the third flow in this cluster of "top" exporters, which corresponds to the decreasing flow of exports from *Partner1* to the new member state CC1, with the flows we highlighted in the other clusters. We see that the same importing country appears in the two *Highly Increasing* singleton clusters, as well as in one flow in the Increasing cluster. Figure 11 shows the three flows of exports from Partner3, Partner4, and Partner5 towards the new member state

CC1.

By comparing Figure 9 with Figure 11, we see that on the one hand CC1 imports from *Partner1* are decreasing, but on the other hand, simultaneously, its imports from the other three partners (with no activity registered before) are increasing. This situation suggests than the activity of these partner countries should be further examined by anti-fraud experts with a view to verifying, first, whether their exports are subject to favorable EU duty rates, and second, the real origin of the trading products. The potential fraud scheme could be the one shown in Figure 12, that is, some of the trading products instead of going directly

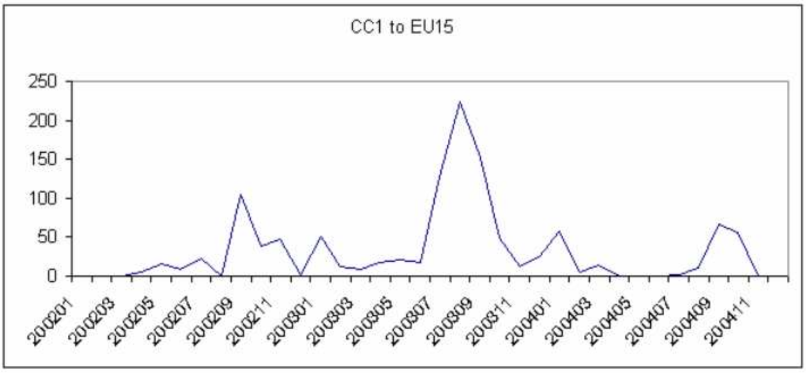

Figure 10: Trade flow from CC1 to EU15

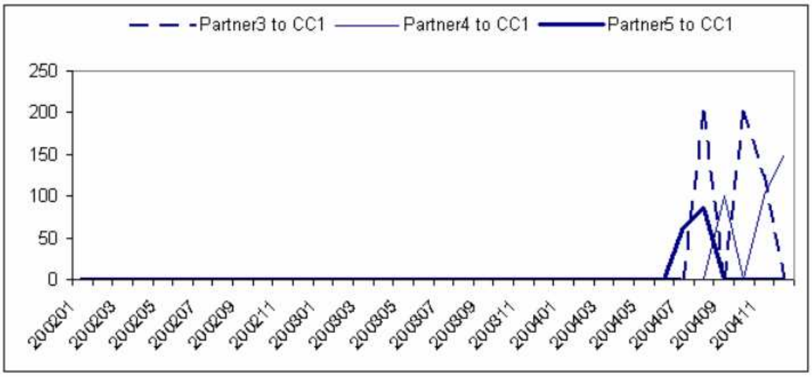

Figure 11: Trade flows from *Partners* 3,4,5 to new member state CC1

into the new member state CC1 (dotted line), they are deviated through the other three countries. In fact, the results of this specific case have been brought to the attention of the European anti-fraud office (OLAF) for further investigation.

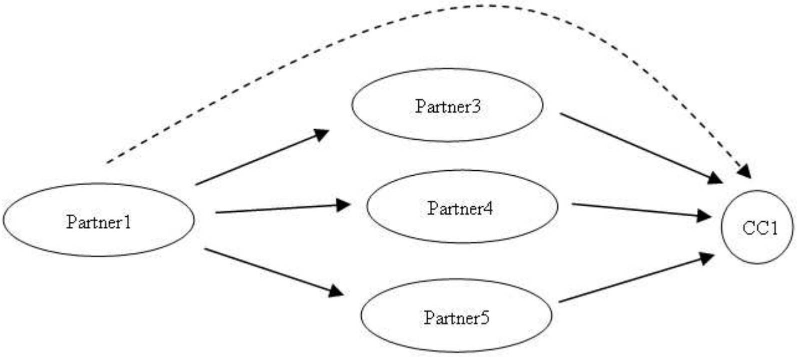

Figure 12: The Potential fraud scheme

6. Conclusion

In this paper we presented a technique for extracting potential fraud schemes from trade data series. Our work was motivated by fraud detection in external trade, as well as other applications that could benefit from efficient change detection and extraction mechanisms applied to time series. We implemented time series forecast models which allowed us to define and detect significant changes through the use of warning sign models. Based on three attributes, namely the sample mean of the historical data, the sample mean of the warning signs, and their ratio, we were able to cluster the trade time series into a number of meaningful classes which lend themselves to the identification of fraudulent patterns. We applied the method to a real case of anti-fraud investigation and the results implied that, depending on the kind of the available data, our approach can come up with highly valuable information for the fraud schemes under consideration. They also pointed to the potential of use of our technique as a building block for a sophisticated fraud extraction system based on data and statistical knowledge guided reasoning. Finally, these results were actually forwarded to the appropriate anti-fraud authorities for further verification.

REFERENCES

[1] Artis M., M. Ayuso, and M. Guillen Detection of automobile insurance fraud with discrete choice models and misclassified claims, *Journal of risk and insurance* **69(3)** (2002), 325–340.

[2] Artis M., M. Ayuso, and M. Guillen Modeling different types of automobile insurance fraud behaviour in the Spanish Market, In *Insurance: Mathematics and Economics* **24** (1998), 67–81.

[3] Barao M.I. and J.A. Tawn Extremal analysis of short series with outliers: sea-levels and athletics records, *Applied Statistics* **48** (1999), 469–487.

[4] Bolton R.J. and D.J. Hand Statistical fraud detection: a review, *Statistical Science* **17** (2002), 235–255.

[5] Brockwell P.J. and R.A. Davis Introduction to Time Series and Forecasting, Springer, New York, 1996.

[6] T. Calderon and B. Green The Use of Neural Networks as an Audit Tool in Fraud Risk Assessment, *Internal Auditing* **16(6)** (2001).

[7] Green B. and J. Choi Assessing the Risk of Management Fraud through Neural Network Technology, *Auditing* **16(1)** (1997), 14–28.

[8] Fanning K., K. Cogger, and R. Srivastava Detection of Management Fraud: A Neural Network Approach, *International Journal of Intelligent Systems in Accounting, Finance and Management* **4** (1995), 113–126.

[9] Kaufman L. and P.J. Rousseeuw Finding Groups in Data, John Wiley and Sons, New York, 1990.

[10] Noncheva V. and C. Moussas A statistical approach to fraud detection in external trade, in: Proceedings of the IADIS International Conference "Applied Computing 2005", Algarve, Portugal, 22-25 Feb., Vol II, Edited by N. Guimaraes and P. Isaias, ISBN: 972-99353-6-X, (2005), 195–200.

[11] R Development Core Team R: A language and environment for statistical computing, `http://www.R-project.org`, 2004.

[12] *Report of the European Anti-Fraud Office (OLAF), Forth acting report for the year ending June 2003, Internet on-line document,* `http://europa.eu.int/comm/dgs/olaf`, 2003.

[13] Robinson M.E. and J.A. Tawn Statistics for exceptional athletics records, *Applied Statistics* **44(4)** (1995), 499–511.

[14] Rousseeuw P.J. Silhouettes: A Graphical Aid to the interpretation and Validation of Cluster Analysis, *Journal of Computational and Applied Mathematics* **20** (1987), 53–65.

[15] Sakamoto Y., M. Ishiguro, and G. Kitagawa Akaike Information Criterion Statistics, Reidel Publishing Company, Dordrecht, 1986.

Charalambos Moussas
Institute for the Protection and Security of the Citizen (IPSC),
Joint Research Centre,
European Commission
Ispra(VA), 21020, Italy
e-mail: charalambos.moussas@jrc.it

Veska Noncheva
Faculty of Mathematics and Informatics,
University of Plovdiv,
24 Tzar Assen Str. 4000 Plovdiv, Bulgaria
e-mail: wesnon@pu.acad.bg