Modelling international trade data with the Tweedie distribution for anti-fraud and policy support

Modelling international trade data with the Tweedie distribution for anti-fraud and policy support

Lucio Barabesi, Andrea Cerasa, Domenico Perrotta, Andrea Cerioli

PII: S0377-2217(15)00798-5
DOI: 10.1016/j.ejor.2015.08.042
Reference: EOR 13193

To appear in: *European Journal of Operational Research*

Received date: 18 February 2015
Revised date: 22 July 2015
Accepted date: 30 August 2015

Please cite this article as: Lucio Barabesi, Andrea Cerasa, Domenico Perrotta, Andrea Cerioli, Modelling international trade data with the Tweedie distribution for anti-fraud and policy support, *European Journal of Operational Research* (2015), doi: 10.1016/j.ejor.2015.08.042

Highlights

- We adapt the Tweedie distribution for modelling economic transactions.
- We address statistical and computational issues in parameter estimation.
- We develop an efficient exact algorithm for random variate generation.
- We empirically show the potential of the Tweedie model for anti-fraud analysis.

Modelling international trade data with the Tweedie distribution for anti-fraud and policy support

Lucio Barabesi[a], Andrea Cerasa[b], Domenico Perrotta[b], Andrea Cerioli[c,*]

[a] *Department of Economics and Statistics, University of Siena; Piazza S.Francesco 7, Siena, Italy*

[b] *European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen; Ispra, Italy*

[c] *Department of Economics, University of Parma; Via Kennedy 6, Parma, Italy*

Abstract

This paper shows the potential of the Tweedie distribution in the analysis of international trade data. The availability of a flexible model for describing traded quantities is important for several reasons. First, it can provide direct support to policy makers. Second, it allows the assessment of the statistical performance of anti-fraud tools on a large number of data sets artificially generated with known statistical properties, which must comply with real world scenarios. We see the advantages of adopting the Tweedie model in several data sets which are particularly relevant in the anti-fraud context and which show non-trivial features. We also provide a systematic outline of the genesis of the Tweedie distribution and we address a number of relevant statistical and computational issues, such as the development of efficient algorithms both for parameter estimation and for random variate generation.

Keywords: Compound Poisson Distribution, Exponential tilting, International trade, Lévy processes, Tempered stable distribution

*Corresponding author. Tel: +39 0521 902491, Fax: +39 0521 902375.
Email addresses: lucio.barabesi@unisi.it (Lucio Barabesi), andrea.cerasa@jrc.ec.europa.eu (Andrea Cerasa), domenico.perrotta@ec.europa.eu (Domenico Perrotta), andrea.cerioli@unipr.it (Andrea Cerioli)

1. Introduction

The regulatory framework of the European Union (EU) reserves to the EU institutions the responsibility of trade relations between the EU Member States and the non-EU countries (TFEU, 2012). It also gives mandates to both institutions and Member States to counter fraud and protect the financial interests of the Union (Council, 1995). This article is grounded on more than fifteen of work conducted in the anti-fraud context by the Joint Research Centre (JRC) of the European Commission (EC), in collaboration with the European Anti-fraud Office (OLAF), academic partners and relevant authorities in Member States. The ultimate target is the development of sound statistical methods for the analysis of international trade data, in order to detect customs frauds (e.g., under-valuation of import duties), trade-related infringements (e.g., money laundering) and cases of circumvention of EU trade regulations (e.g., anti-dumping and countervailing measures). The resulting signals are used in the definition of audit plans or for initiating investigations.

The statistical tools that the JRC has implemented for anti-fraud analysis involve the inspection of the quantity of specific products which are imported in, or exported from, the EU market. It is crucial to rely on flexible statistical models for describing the distribution of these quantities for a large number of traded products. First, such models will provide direct support to the EU policy makers, in the form of tools for monitoring the effect of policy measures and for deciding how to react against international trade distortions originating in and outside the EU. For example, statistical models of trade can provide factual background for the official communications on trade policy (see, e.g., European Commission, 2012), or for the related preparatory technical documents (Lejeune et al., 2013).

Another important goal for modelling traded quantities is the assessment of the statistical performance of alternative methods used for finding relevant patterns in international trade data. For anti-fraud purposes, outlier detection and robust clustering tools are typically required (see, e.g., Fogelman-Soulie et al., 2008; Riani et al., 2009; Cerioli, 2010; Perrotta and Kopustinskas, 2010; Cerioli and Perrotta, 2014). It is very difficult to derive analytical results for such methods in finite samples, especially when non-Gaussian distributions are involved. Therefore, the methods need to be compared, evaluated and eventually tuned on a large number of data sets artificially generated with known statistical properties, which must reflect

the distributions observed in trade data. The compliance of such models with real world scenarios is of paramount importance, especially in the case of legal disputes, when statistical evidence is evaluated in Court. For example, Cerioli and Perrotta (2014) tackle the problem of assessing alternative robust regression clustering tools in a particular anti-fraud problem. They thus assume a Gamma distribution for simulating traded quantities, but with a rather ad hoc and product-specific choice of parameter values. Their approach, although effective in specific applications, is thus difficult to replicate in large-scale analyses of different products and lacks general probabilistic motivation. The adoption of nonparametric tools for density estimation is also problematic for similar reasons, as it does not allow for ready simulation of artificial data sets. Furthermore, it is not straightforward how to extend standard density estimation methods to the case of distributions with a non-negligible mass at (or close to) zero, as it often happens in the case of trade data.

This article concentrates on the Tweedie distribution, a flexible three-parameter model which is fitted to the traded quantity of products of major importance for anti-fraud purposes and trade policy support. Similar models have been successfully adopted in different application fields, and seem to be particularly attractive in economics and finance (Menn and Rachev, 2005; Rachev et al., 2011; Francq and Zakoian, 2013; Babaei et al., 2015). The main applied contribution of our work is to show that the model that we propose is able to capture most of the basic features observed in international trade data. Such features are not easy to analyze, due to the combination of economic activities and normative constraints. Indeed, we typically have to face with markedly skew empirical distributions with heavy tails, a large number of rounding errors in small-scale transactions due to data registration problems, and structural zeros arising because of confidentiality issues related to national regulations. We see the advantages of adopting the Tweedie distribution in several data sets which are particularly relevant for the tasks sketched above. One is the category of the *Petroleum oils and oils obtained from bituminous minerals, crude*, of great importance for governments and policy makers. The second application area includes a range of *fraud-sensitive products* that are regularly monitored by the anti-fraud partners of the JRC.

To achieve our applied goal we also face a number of methodological and computational issues. First, we provide a systematic introduction to the Tweedie distribution, unifying contributions which are currently scattered in the literature. We emphasize the stochastic genesis of this model either

as an exponentially tilted stable or a compound Poisson distribution. We also describe how the Tweedie distribution can be embedded in the theory of Lévy processes, which has direct relevance to international trade. We argue that our systematic outline could be easily understood and provide motivation to applied scientists analyzing trade data. Then, we suggest new efficient computational algorithms both for estimating the parameters of the Tweedie model from observed trade data and for random variate generation in Monte Carlo studies. Computational performance of the estimation method represents a crucial issue in anti-fraud problems, as the model needs to be estimated on thousands of possible products. Similarly, the availability of efficient and easy-to-implement simulation algorithms is an important ingredient for performing large scale assessments of the effectiveness of statistical anti-fraud tools.

The rest of the paper is organized as follows. In order to sharpen the focus of our work, in §2 we make some connections with approaches in Operations Research and Economic Theory dealing with international trade. The genesis and format of the trade data that we need to model are introduced in §3. Then, §4 gives an account of the Tweedie distribution, relates it to Lévy processes and motivates its use for trade data. In §5 we propose an efficient approach to parameter estimation and we provide efficient random variate generators from the Tweedie model. In §6 we apply the model to empirical cases of major importance in anti-fraud analysis, while §7 provides some comparisons with respect to competing models. In §8 we assess the stability of our models over time. Closing remarks are given in §9. Appendix A contains the pseudo-code of our simulation algorithms, while Appendix B reports descriptive statistics for the trade data that we analyze.

2. Trade models in Operations Research and related fields

The dynamics of international trade has been traditionally studied on the basis of economic theory models using methodologies from different disciplines, in particular Operations Research, Information Theory and entropy optimization principles. A classic starting point is a linear programming cost minimization model that relates the flows of commodities between two geographic regions to the production, consumption and costs of transportation (see, e.g., Harris, 1974). In this model, if q_{ij} is the quantity of product shipped from an origin i to a destination j and c_{ij} is the related transportation, eligible flows configurations are found by minimizing the objective

function

$$\sum_i \sum_j c_{ij} q_{ij},$$

subject to the constraints $q_{ij} \geq 0$, $\sum_j q_{ij} = Q_{i\cdot}$ for each origin i, where $Q_{i\cdot}$ is the total production at origin i (used as surrogate for the total export from i), and $\sum_i q_{ij} = Q_{\cdot j}$ for each destination j, where $Q_{\cdot j}$ is the total consumption at destination j (used as surrogate for the total import at destination j). Additional constraints are often considered to address realistic trade patterns, a popular one being the entropy function $-\sum_i \sum_j q_{ij} \ln q_{ij}$ (see e.g. Erlander, 1977, 1982).

A second research line applies entropy maximizing principles to various sets of assumptions of typical trade frameworks, to derive the most unbiased (or less informative) distributions for the trade quantities or for other trade statistics of interest. This is achieved by specifying a finite number of moment values and by choosing, out of all the probability distributions with these moment values, the one which maximizes an entropy measure. In the discrete case this is the Shannon entropy or (in case a prior probability distribution is also given) the Bayesian entropy. Lagrange multipliers are then used to solve the constrained problem. The principle, originally proposed for discrete data, was generalized to the continuous case by Jaynes (1968). Its application to modelling transportation and international trade are described in Kapur (1993, pp. 425–428). Other examples of use of the Maximum Entropy principle in international trade are discussed by Wilson (1970). The approach is flexible and can be also combined with rather complex assumptions on trade dynamics; see, e.g., the integration of Lotka-Volterra "predator–prey" type dynamics recently proposed by Fray and Wilson (2012), or the possibility to include transaction costs (Zhang et al., 2012).

To give an idea of the distributions that can be derived for modelling international trade with the Maximum Entropy principle let us now assume, following Kapur (1993), that $Q_{\cdot j}$ is the total import at destination j in the current year; similarly, $Q_{i\cdot}$ is the total export at origin i in the current year. In addition, let $\sum_i \sum_j q_{ij} = Q$ and define $p_{ij} = q_{ij}/Q$, $p_{i\cdot} = Q_{i\cdot}/Q$ and $p_{\cdot j} = Q_{\cdot j}/Q$. We do not know the q_{ij} values for the current year, but we can assume to know averages from previous years and define apriori estimates for the unknown proportions p_{ij}, say p^*_{ij}. In this setting, by maximizing the entropy function under the given constraints on past averages and on current total import/export values obtained from an external source (e.g.

information on production/consumption, as above), we find the maximum entropy estimates

$$\tilde{p}_{ij} = p^*_{ij} p_{i\cdot} p_{\cdot j} a_i b_j,$$

where a_i and b_j are determined by using $a_i \sum p^*_{ij} p_{\cdot j} b_j = 1$ and $b_j \sum p^*_{ij} p_{i\cdot} a_i = 1$. The maximum entropy estimates $\tilde{p}_{ij}$ will be finally used to infer the volumes q_{ij} traded in the current year between each pair of countries.

Although rooted in the same types of data, these research lines are different from our approach. Typically the goal is to build input-output models for the trading countries. For example, one can use the model to check if a certain policy objective, e.g. to increase GDP per capita, is feasible and what is the best trade setting to achieve it. Then, the trade data are used to validate the relationships between the model factors, calibrate the model parameters and study the dependence of the initial conditions on the model dynamics. The tools of sensitivity analysis can be an extremely useful support for such a purpose (Saltelli et al., 2004).

On the other hand, the aims of our statistical perspective are to justify on solid grounds the choice of a certain density function for the trade data population and to fit its parameters on the basis of the observed trade sample data. Our selected model, i.e. the Tweedie distribution, has a clear interpretation according to the actual value of its estimated parameter and also a substantive motivation in terms of the generating economic process, as shown in §§4.1–4.3 below. Furthermore, we can also verify the quality and stability of our estimated models by checking their distribution for data chosen according to different criteria (randomly, in different time windows, etc.).

To understand the main links between the two perspectives we recommend reading the excellent book of Kapur (1993, pp. 266–291), which also contains elegant mathematical programming formulations of relevant entropy maximization problems (pp. 553–564) and interesting links between the probability distributions derived through the maximum entropy approach and the frequencies in the cells of contingency tables (pp. 252–265). In particular, it shows how to derive the classical χ^2 test for testing the independence of attributes in contingency tables from the maximum entropy principle. This and other relationships were brought up by Theil (1967) and, again in relation to international trade, also by Wilson (1970).

Table 1: A glimpse into the data set for product "Petroleum oils and oils obtained from bituminous minerals, crude" (CN-27090090).

Import country	Export country	CN code	Month	Traded quantity
EU_1	$\overline{EU}_1$	27090090	August 2008	1,401,221.0
EU_1	$\overline{EU}_2$	27090090	August 2008	146,360.1
EU_3	$\overline{EU}_2$	27090090	August 2008	17,422.0
...	...	...	...	...

3. Trade data and the related statistical challenges

Our data come from the official extra-EU trade statistics, *Extrastat*, extracted from the COMEXT database of Eurostat. We consider monthly aggregates of trade quantities for each product, country of origin and country of destination, registered in the period from August 2008 to July 2012. We construct one data set for each product. A glimpse into one of them is provided in Table 1, where $EU_1, EU_2, \ldots$ denote the EU Member State and $\overline{EU}_1, \overline{EU}_2, \ldots$ are non-EU countries. Each observation in the data set is the amount of trade for the selected product that took place from a non-EU country to a Member State in a given month. This observation may be the result of a single transaction, or, more frequently, it is the aggregate over several transactions possibly involving different traders. The aggregates are built from the customs declarations collected from individuals or companies by the Member States, following a strictly regulated process (Eurostat, 2006). The products are specified according to a numeric code of the Combined Nomenclature (CN). We model data specified at the maximum level of accuracy generally accessible for both imports and exports, that is 8 CN digits. This level of classification, containing more than 10,000 sub-headings, in general is sufficiently detailed to distinguish the products by their material, function and degree of processing.

Depending on the product, the traded quantities are expressed in tons of net mass and/or in supplementary units (liters, number of items, etc.). The weight of packaging is not included in the net mass. With the aim of simplifying customs operations, small scale transactions may not be declared by the traders. The thresholds are fixed by Member States within the maximum limit of 1 ton for the net masses and 1000 euros for the values. However, the national administrations have to make estimations for trade below the thresholds, for which there are no common methodologies. In addition, when

statistical units risk to disclose information on individual traders, the declared quantities and/or values may be hidden to the user by replacing them with a zero. Again, each Member State fixes the criteria to decide which data units should be treated as confidential. As a result of these issues, the data include monthly aggregates with quantity rounded to zero but of positive value and also (but less frequently) aggregates of positive quantity but value rounded to zero. Data quality is quite heterogeneous across countries and products and, for this reason, is subject to constant monitoring by the statistical authorities and customs services. However, only macroscopic outliers are removed or corrected.

The typical combination of trading factors, regulation constraints and reporting errors (either fraudulent or not) gives rise to empirical distributions of quantities that are markedly skew, with heavy tails and a considerable number of transactions very close, or even equal, to zero. To be concrete, Figure 1 reports the empirical distribution of traded quantities, in millions of tons, for product "Petroleum oils and oils obtained from bituminous minerals, crude" (CN-27090090) in the selected time frame (August 2008 to July 2012, chosen to cover a sufficiently long period). This product is the most relevant one for the EU balance sheet, covering alone around 17% of all EU trade imports. Its distribution, which is described and analyzed in detail in §6.1, is the result of 6,651 transactions, concerning imports of the selected product in all Member States from non-EU countries. With a slight graphical abuse, the black bin of the histogram depicted in the left-hand panel corresponds to 286 transactions for which the recorded quantity is *exactly zero*, due to rounding errors or – more often – to confidentiality issues. The long and slowly decaying right-tail is paramount, as is the spike of transactions for which the recorded quantity is positive but very close to zero. Indeed, in this example slightly more than 20% of the observations fall in the first strictly positive bin of the histogram, which corresponds to a traded quantity of about 45,000 tons. The largest transaction exceeds 4,220,000 tons. The empirical masses concentrated both at the origin and in a relatively small neighbourhood of it have great influence on parameter estimation for the candidate statistical models which are fitted to such data, as well as on the assessment of model accuracy. Any credible modelling must deal carefully and appropriately with the small values, since they comprise a large part of the data set. We thus need a statistical model with enough flexibility to accommodate both these masses and the tail of the distribution.

We emphasize that, although the qualitative tendency observed in inter-

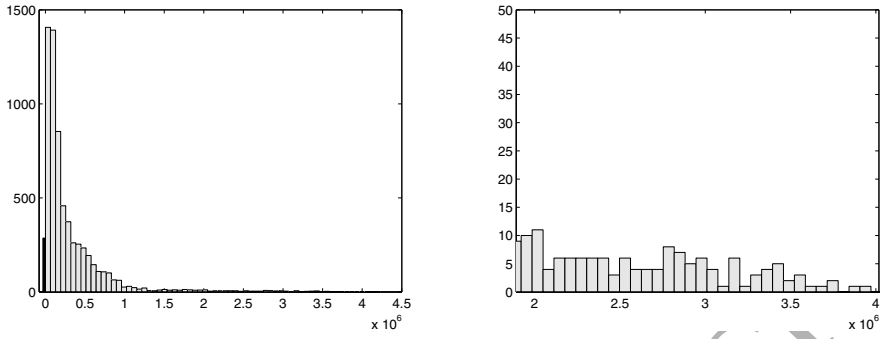


Figure 1: Left: empirical distribution of traded quantities, in millions of tons, for product "Petroleum oils and oils obtained from bituminous minerals, crude" (CN-27090090), recorded in COMEXT from August 2008 to July 2012. In this example 4.3% of the data are null (black bin of the histogram). Right: zoom into the right-hand tail.

national trade data is general, the quantitative precise specification of the features described above is product-dependent and cannot be anticipated without visual inspection of the data, which is impossible in routine anti-fraud applications. Therefore, our distributional model must fit adequately well to a large number of data sets without prior specification of any tuning constant. It is also impractical to model the rounding and recording errors precisely, given the currently available information, since regulations and controls still vary considerably from Member State to Member State. We find that our model based on the Tweedie distribution has the required flexibility for many products which are relevant in anti-fraud analysis.

4. Genesis of the Tweedie distribution

The Tweedie distribution has been popularized and analyzed at length by Jörgensen (1987), following the seminal ideas of Tweedie (1984). For some parameter values the distribution has been also introduced by Hougaard (1986). Further discussion is contained in Aalen (1992) and Barndorff-Nielsen and Shephard (2001). Hougaard (1986) gives the Laplace transform of the Tweedie r.v. X with the following parametrization:

$$L_X(s) = \mathrm{E}[\exp(-sX)] = \exp[(\delta/\alpha)(\theta^\alpha - (\theta+s)^\alpha)], \qquad \mathrm{Re}(s) > 0, \quad (1)$$

where the parameter space is given by

$$(\alpha, \theta, \delta) \in \{]-\infty, 1]\times]0, \infty[\times]0, \infty[\} \cup \{]0, 1] \times \{0\}\times]0, \infty[\}$$

and the boundary for $\alpha = 0$ has been managed for analytical continuity.

It is worth considering the family morphology as parameters vary. Since α crucially determines the characteristics of the Tweedie distribution, we will focus on the two fundamental subsets of its parameter space in §4.1 and §4.2, respectively.

4.1. The Tweedie distribution as an exponentially-tilted stable distribution

We first assume that $\alpha \in]0, 1]$. Let Z be a positive Stable r.v. of index α (see Sato, 1999, for details). The Laplace transform of the scaled r.v. $Y = (\delta/\alpha)^{1/\alpha} Z$ is

$$L_Y(s) = \exp(-(\delta/\alpha)s^\alpha), \qquad \mathrm{Re}(s) > 0.$$

In such a case, expression (1) may be rewritten as

$$L_X(s) = \frac{\exp[-(\delta/\alpha)(\theta+s)^\alpha]}{\exp(-\delta\theta^\alpha/\alpha)} = \frac{L_Y(\theta+s)}{L_Y(\theta)},$$

i.e. an exponentially-tilted stable r.v. X with tilting parameter θ is achieved. See, e.g., Devroye and James (2014), Lijoi and Prunster (2014), and Favaro and Nipoti (2014) for details on this distribution.

If f_X and f_Y are the p.d.f.'s of X and Y with respect to the Lebesgue measure on $\mathbb{R}$, we obtain

$$f_X(x) = \frac{\exp(-\theta x) f_Y(x)}{L_Y(\theta)}.$$

Since

$$f_Y(x) = \sum_{n=1}^{\infty} \frac{(-1)^n \delta^n}{n!\Gamma(-n\alpha)\alpha^n}\, x^{-n\alpha-1} I_{[0,\infty[}(x),$$

where I_B is the usual indicator function of a given set B (Sato, 1999), it follows

$$f_X(x) = \exp(-\theta x + \delta\theta^\alpha/\alpha) \sum_{n=1}^{\infty} \frac{(-1)^n \delta^n}{n!\Gamma(-n\alpha)\alpha^n}\, x^{-n\alpha-1} I_{[0,\infty[}(x).$$

A number of special distributions are contained in the family for $\alpha \in]0, 1]$: the positive Stable r.v. Z is achieved for $\theta = 0$ and $\delta = \alpha$, while $\alpha = 1/2$ yields the inverse Gaussian distribution. In addition, the Dirac mass at

$x = \delta$ is achieved if $\alpha = 1$ for any $\theta \geq 0$. Finally, the Gamma distribution with shape parameter δ and scale parameter $(1/\theta)$ is obtained as $\alpha \downarrow 0$. The parameters α and θ thus rule the tail heaviness of the distribution. Specifically, for $\alpha \in]0, 1]$, in the limiting case where $\theta = 0$ moments exist solely for order $r < \alpha$, while for $\theta > 0$ the moments of all orders exist, even if the right tail tends to be heavier as θ approaches zero.

4.2. The Tweedie distribution as a compound Poisson distribution

When $\alpha \in]-\infty, 0[$, expression (1) may be rewritten as

$$L_X(s) = \exp[-(\delta\theta^\alpha/\alpha)((1 + s/\theta)^\alpha - 1)], \tag{2}$$

i.e. the law of X may be expressed as a compound of a Poisson distribution, with parameter $(-\delta\theta^\alpha/\alpha)$, and Gamma r.v.'s with shape parameter $(-\alpha)$ and scale parameter $1/\theta$ (Aalen, 1992). If $\{G_n\}_{n\geq 1}$ is a sequence of copies of such Gamma r.v.'s, while N is a Poisson r.v. with parameter $(-\delta\theta^\alpha/\alpha)$, X is thus stochastically represented as

$$X \overset{\mathcal{L}}{=} \sum_{n=1}^{N} G_n, \tag{3}$$

with the assumption that X degenerates at zero if $N = 0$. Hence, in this case X displays a mixed distribution, given by a convex combination of a Dirac distribution (with mass at zero) and an absolutely continuous distribution. By exploiting the properties of Gamma r.v.'s, the distribution function of X is

$$F_X(x) = \exp(\delta\theta^\alpha/\alpha) I_{[0,\infty[}(x) + [1 - \exp(\delta\theta^\alpha/\alpha)] \int_{-\infty}^{x} g_X(u)\mathrm{d}u,$$

where g_X is the p.d.f. of X conditioned to the event $\{N > 0\}$. From (2), we obtain

$$g_X(x) = \frac{\exp(-\theta x + \delta\theta^\alpha/\alpha)}{1 - \exp(\delta\theta^\alpha/\alpha)} \sum_{n=1}^{\infty} \frac{(-1)^n \delta^n}{n!\Gamma(-n\alpha)\alpha^n} x^{-n\alpha-1} I_{[0,\infty[}(x).$$

Hence, g_X coincides with f_X for $\alpha \in]0, 1]$ up to the factor $[1 - \exp(\delta\theta^\alpha/\alpha)]^{-1}$.

Again, some special distributions are contained in the family: the non-central Gamma distribution of zero shape is achieved for $\alpha = -1$, while the Poisson distribution is obtained as $\alpha \to -\infty$. In turn, the Gamma distribution with shape parameter δ and scale parameter $1/\theta$ is accomplished as $\alpha \uparrow 0$.

4.3. The Tweedie distribution embedded in the theory of stochastic processes

In order to motivate the Tweedie model for trade data, it is interesting to provide the genesis of the distribution in the framework of Lévy processes. Informally speaking, a non-negative Lévy process is a stochastic jump process with non-negative, independent, time-homogeneous increments. See Sato (1999) for a formal definition. The Laplace transform of such a process, say X_t with $t \geq 0$, is given by the Lévy-Khintchine representation

$$L_{X_t}(s) = \mathrm{E}[-sX_t] = \exp[-t\psi(s)], \qquad \mathrm{Re}(s) > 0,$$

where t is the time parameter of the process, while the function

$$\psi(s) = \int (1 - e^{-sx})\nu(\mathrm{d}x)$$

is referred to as the characteristic exponent of the Lévy process and ν is the so-called Lévy measure. The family of Lévy processes encompasses many remarkable special cases, such as the compound Poisson processes, the Gamma processes, the Stable processes, among others. Indeed, all non-negative Lévy processes are limits of compound Poisson processes.

In trade, an observation corresponding to a given positive variable may be ideally seen as the result of a cumulative process in which the marginal contributions occur independently during time and at the same rate. Indeed, the COMEXT data introduced in §3 are precisely the result of such an aggregation, involving all the imports of the selected product in a given month. It is apparent that this concept may be modelled as a Lévy process, and the stochastic outcome at a given time is actually the observation on hand. Hence, it is interesting for our purposes to embed the Tweedie distribution in the framework of Lévy processes.

When $\alpha \in]0, 1[$, by taking an exponentially-tilted stable Lévy measure such that

$$\frac{\nu(\mathrm{d}x)}{\mathrm{d}x} = \frac{\lambda}{\Gamma(1-\alpha)}\, x^{-1-\alpha} \exp(-\theta x) I_{[0,\infty[}(x),$$

where $\lambda > 0$, it follows that

$$\psi(s) = \frac{\lambda}{\Gamma(1-\alpha)} \int_0^\infty (1 - e^{-sx}) x^{-1-\alpha} \exp(-\theta x) \mathrm{d}x = \frac{\lambda}{\alpha}((\theta + s)^\alpha - \theta^\alpha).$$

Hence, the Laplace transform of the process X_t at time t is given by

$$L_{X_t}(s) = \exp[(\lambda t/\alpha)(\theta^\alpha - (\theta + s)^\alpha)], \qquad \mathrm{Re}(s) > 0.$$

By letting $\lambda t = \delta$ for a fixed t, the Tweedie r.v. for $\alpha \in]0, 1[$ may thus be considered as the "outcome" of an exponentially-tilted stable Lévy process. On the contrary, a non-stochastic process is actually involved in the case $\alpha = 1$, since the Tweedie distribution degenerates to a Dirac mass in that situation (see Brix, 1999, p. 933).

Subsequently, let us take $\alpha \in]-\infty, 0[$. In this case, a Poisson process of rate λ is running on time scale t and each jump is assumed to be a Gamma r.v. (independent of the past) with shape parameter $(-\alpha)$ and scale parameter $1/\theta$. This compound Poisson process X_t is actually the sum of the Gamma r.v.'s up to time t. The Laplace transform of the process X_t at time t is given by

$$L_{X_t}(s) = \exp[\lambda t((1 + s/\theta)^{\alpha} - 1)], \qquad \mathrm{Re}(s) > 0,$$

and hence X_t is a Lévy process with characteristic exponent given by

$$\psi(s) = \lambda(1 - (1 + s/\theta)^{\alpha}).$$

By letting $\lambda t = -\delta\theta^{\alpha}/\alpha$ for a fixed t, it is apparent that the Tweedie r.v. for $\alpha \in]-\infty, 0[$ may be considered as the "outcome" of this compound Poisson process.

4.4. Potential of the Tweedie model for trade data

It is intuitively apparent from the discussion outlined above that the flexibility of the Tweedie distribution makes it an attractive candidate for modelling trade data sets which follow the qualitative pattern depicted in Figure 1. According to its actual parameter values, it can accommodate tails of different length and thickness, as well as allowing for the height and the position of the spike close to the origin.

We emphasize that the Tweedie model is widely applicable to data from different countries and also of possible different quality concerning under-reporting issues. As this model fits automatically the presence of null and very small values, it does not require that the user specifies the actual threshold for rounding errors, nor the the precise mechanism which generates under-reporting for small trade quantities. These are clearly appealing features for routine application to a wide range of products ranging from oil to textile goods and food. In addition, the contribution of each Member State to trade is often very different for each product, thus making the effect of state-specific regulations difficult to anticipate in general terms.

The relationship with Lévy processes provides an additional motivation for the use of the Tweedie distribution with international trade data. In fact, the quantities recorded in a given temporal and spatial market are obtained as additive aggregates of transactions occurred at a smaller scale, up to the individual company level. This cumulative trading process may be represented by sums of increments of the type $X_{t_i} - X_{t_{i-1}}$, so that X_t becomes the cumulative traded quantity of a given product up to time t. Therefore, modelling the generation mechanism of trade data through the Tweedie distribution may also represent an appealing choice from an economic point of view.

As suggested by one referee, further insights on the substantial mechanism generating trade data can be achieved on the basis of the Lévy process representation through the stochastic measure considered by Brix (1999). Indeed, Brix (1999) emphasizes that this stochastic measure may be seen as a sum of infinitely many terms for $\alpha \in]0,1[$, since the Lévy measure ν does not integrate, i.e. $\nu(]0,\infty[) = \infty$. Hence in this case the Lévy process has infinitely many jumps and 0 is the accumulation point of the infinite sequence of jumps. On the other hand, for $\alpha \in]-\infty,0[$ the stochastic measure is a sum of a finite number of terms, in such a way that the jumps follow a Gamma distribution with shape parameter $(-\alpha)$ and scale parameter $1/\theta$. The two different features may thus be taken to represent alternative trading schemes suitable for different markets. In fact, for many consumer goods it may be reasonable to assume that the traded quantities arise from a large number of transactions that take place almost instantaneously, as it happens in the case $\alpha \in]0,1[$. On the other hand, for products like the Petroleum oils introduced in §3, the traded quantities are typically obtained through a limited number of relatively large transactions, corresponding to a finite number of jumps in the underlying stochastic mechanism, as is the case when $\alpha \in]-\infty,0[$. These jumps might be potentially very high, but their effect is tempered by θ. The estimated values of α and θ in the Tweedie model can thus shed light on the actual trade process that rules the observed quantities. Examples of this interpretation are shown in §6.

5. Statistical and computational issues

5.1. Parameter estimation

If $X_1, X_2, \ldots, X_n$ represent n independent copies of a Tweedie r.v. X with $\alpha \in]0,1]$, the likelihood function (with respect to the Lebesgue measure) is

given by

$$L(\alpha, \theta, \delta) = \exp(n\delta\theta^{\alpha}/\alpha) \prod_{i=1}^{n} \exp(-\theta x_i) \sum_{j=1}^{\infty} \frac{(-1)^j \delta^j x_i^{-j\alpha-1}}{j!\Gamma(-j\alpha)\alpha^j}. \tag{4}$$

In contrast, when $\alpha \in]-\infty, 0[$, M observations out of n are null, where M is a Binomial r.v. with parameters n and $\exp(\delta\theta^{\alpha}/\alpha)$. By reindexing in such a way that the first $(n-m)$ sample realizations are non-null, the resulting likelihood function (with respect to the Dirac measure at zero and the Lebesgue measure) is given by

$$L(\alpha, \theta, \delta) = \exp(n\delta\theta^{\alpha}/\alpha) \prod_{i=1}^{n-m} \exp(-\theta x_i) \sum_{j=1}^{\infty} \frac{(-1)^j \delta^j x_i^{-j\alpha-1}}{j!\Gamma(-j\alpha)\alpha^j}. \tag{5}$$

Obviously, the series in (4) and (5) should be truncated at a given value in order to practically handle the function. We may expect the resulting approximation to be accurate with few terms, at least when the observed x_i are not too small or too large (Dunn and Smyth, 2005). Alternatively, if the series does not converge quickly, the suggestions provided by Palmer et al. (2008) and Dunn and Smyth (2008) for numerical inversion of the Laplace transform may be considered.

In the analysis of international trade, data sets often contain thousands of observations. Furthermore, hundreds or thousands of data sets typically need to be scrutinized in sequence, both for routine inspection over different categories of products and for performance assessment of anti-fraud tools. The computational performance of estimation methods thus becomes a crucial issue. In order to set up a computationally efficient approach to the evaluation of the likelihood functions (4) and (5), we consider the generalization of the Inversion Theorem given by Barabesi and Pratelli (2015).

First, let us assume that $\alpha \in]0, 1]$. In this case, if h is a positive measurable function defined on $\mathbb{R}$ such that $0 < \mathrm{E}[h(X)] < \infty$ and provided that $\psi_{h,X}(t) = \mathrm{E}[h(X)\exp(\mathrm{i}tX)]$, where i is the imaginary unit, the expression

$$f_X(x) = \frac{1}{2\pi h(x)} \int_{-\infty}^{\infty} \exp(-\mathrm{i}tx)\psi_{h,X}(t)\mathrm{d}t \tag{6}$$

holds a.e. with respect to the Lebesgue measure on $\mathbb{R}$. The integral in (6) eventually represents a Cauchy principal value. In this case, if $\theta \neq 0$ and by

choosing $h(x) = \exp(cx)$ with $c < \theta$, this representation reduces to

$$f_X(x) = \frac{\exp(-cx + \delta\theta^\alpha/\alpha)}{2\pi} \int_{-\infty}^{\infty} \exp[-\mathrm{i}tx - (\delta/\alpha)(\theta - c - \mathrm{i}t)^\alpha]\,\mathrm{d}t. \quad (7)$$

After a little algebra and a change of variable in the integral, expression (7) gives rise to

$$f_X(x) = \frac{1}{2\pi\delta^{1/\alpha}} \exp[-(\delta^{1/\alpha}\theta - \phi)(\delta^{-1/\alpha}x) + \delta\theta^\alpha/\alpha]\, \eta_{\alpha,\phi}(\delta^{-1/\alpha}x), \quad (8)$$

where $\phi = \delta^{1/\alpha}(\theta - c)$, while

$$\eta_{\alpha,\phi}(u) = \int_{-\infty}^{\infty} \exp[-\mathrm{i}tu - (1/\alpha)(\phi - \mathrm{i}t)^\alpha]\mathrm{d}t. \quad (9)$$

If the real and the imaginary parts of the integrand are considered, by means of a change of variable, expression (9) may be rewritten as

$$\eta_{\alpha,\phi}(u) = 2 \int_0^{\infty} \exp[-k_{\alpha,\phi}(t)\cos(l_{\alpha,\phi}(t))]\cos[-tu + k_{\alpha,\phi}(t)\sin(l_{\alpha,\phi}(t))]\,\mathrm{d}t, \quad (10)$$

where

$$k_{\alpha,\phi}(t) = \alpha^{-1}(\phi^2 + t^2)^{\alpha/2} \text{ and } l_{\alpha,\phi}(t) = \alpha \arctan(t/\phi). \quad (11)$$

In practice, the combination of (8) and (10) allows us to rephrase f_X in a new integral form, in such a way that the integral solely depends on the parameter α, since ϕ may be pre-fixed at a given value – obviously, by choosing $c = \theta - \delta^{-1/\alpha}\phi$. It is clear that ϕ should be selected as a convenient value for the evaluation of the integral. As an example, $\phi = 1$ could be a compromise choice, since this selection implicitly leads us to consider a Tweedie r.v. whose expectation equals to one in the integral part. In addition, f_X is expressed in such a way that $\gamma = \delta^{1/\alpha}$ appears as the "natural" scale parameter. Hence, the likelihood function may be rewritten as

$$L(\alpha, \theta, \delta) = \frac{\exp(n\delta\theta^\alpha/\alpha)}{(2\pi\delta^{1/\alpha})^n} \prod_{i=1}^{n} \exp[-(\delta^{1/\alpha}\theta - \phi)(\delta^{-1/\alpha}x_i)]\, \eta_{\alpha,\phi}(\delta^{-1/\alpha}x_i). \quad (12)$$

Finally, the case $\theta = 0$ – corresponding to the positive Stable distribution – may be handled by means of suitable methods existing in the literature.

When $\alpha \in]-\infty, 0[$, it is not possible to adopt immediately the generalization of the Inversion Theorem, since in this case the Tweedie r.v. displays a mixed distribution. However, it is feasible to apply the theorem to g_X. By remarking that the Laplace transform of the r.v. X conditioned on the event $\{N > 0\}$ is given by $(L_X(s) - \exp(\delta\theta^\alpha/\alpha))/(1 - \exp(\delta\theta^\alpha/\alpha))$, g_X may expressed as

$$g_X(x) = \frac{1}{2\pi\delta^{1/\alpha}} \frac{\exp[-(\delta^{1/\alpha}\theta - \phi)(\delta^{-1/\alpha}x) + \delta\theta^\alpha/\alpha]}{1 - \exp(\delta\theta^\alpha/\alpha)} \varphi_{\alpha,\phi}(\delta^{-1/\alpha}x), \tag{13}$$

where

$$\varphi_{\alpha,\phi}(u) = \int_{-\infty}^{\infty} \exp(-\mathrm{i}tu)[\exp(-(1/\alpha)(\phi - \mathrm{i}t)^\alpha) - 1]\mathrm{d}t. \tag{14}$$

In this case, (14) may be reformulated as

$$\varphi_{\alpha,\phi}(u) = 2 \int_0^{\infty} (\exp[-k_{\alpha,\phi}(t)\cos(l_{\alpha,\phi}(t))]\cos[-tu + k_{\alpha,\phi}(t)\sin(l_{\alpha,\phi}(t))] - \cos(tu))\mathrm{d}t.$$

Therefore, the likelihood function reduces to

$$L(\alpha, \theta, \delta) = \frac{\exp(n\delta\theta^\alpha/\alpha)}{(2\pi\delta^{1/\alpha})^{n-m}} \prod_{i=1}^{n-m} \exp[-(\delta^{1/\alpha}\theta - \phi)(\delta^{-1/\alpha}x_i)]\, \varphi_{\alpha,\phi}(\delta^{-1/\alpha}x_i). \tag{15}$$

We finally remark that the integral representations (8) and (13) also provide the basis for finding suitable expressions of the information matrix and its estimated counterpart.

5.2. Random variate generation

We now discuss alternative ways of generating random values from the Tweedie distribution. This task is an important ingredient in any Monte Carlo study which aims at assessing the statistical properties of anti-fraud tools (Cerioli and Perrotta, 2014), when applied to products for which the Tweedie distribution represents a satisfactory model.

We start with the case $\alpha \in]0, 1[$, for which Hofert (2011) suggests both a naive algorithm and an improved version of it, even if these proposals are inefficient as δ increases. On the other hand, Devroye (2009) introduces a more efficient – even if complex to implement – algorithm, subsequently corrected and improved by Hofert (2011). Finally, Ridout (2009) considers

Table 2: Rejection constant $A(\rho_1^*, \rho_2^*)$ for Algorithm 1 and, in parenthesis, for the algorithm of Devroye (2009).

			α	0.1	0.3	0.5	0.7	0.9
δ	1	θ	1	2.33(7.32)	1.78(7.06)	1.72(6.70)	1.96(5.96)	3.47(4.25)
			5	1.93(4.03)	1.43(3.91)	1.30(3.93)	1.27(7.33)	1.46(6.50)
			10	1.82(3.91)	1.36(3.53)	1.25(3.30)	1.21(3.40)	1.24(7.20)
δ	5	θ	1	1.21(2.08)	1.20(2.21)	1.20(2.56)	1.21(3.41)	1.38(6.71)
			5	1.19(2.04)	1.17(2.02)	1.16(2.03)	1.15(2.07)	1.16(2.80)
			10	1.19(2.03)	1.16(2.00)	1.15(1.99)	1.14(1.99)	1.15(2.13)

numerical inversion of the Laplace transform. See also Bianchi and Fabozzi (2014).

Barabesi and Pratelli (2015) propose a universal algorithm which combines efficiency and simplicity if applied to the Tweedie r.v. for $\alpha \in]0, 1[$ (see also Barabesi and Pratelli, 2014). Following their approach, let us consider the function

$$a(\rho) = \frac{\exp(\delta\theta^\alpha/\alpha)}{\pi} \int_0^\infty \exp[-\delta k_{\alpha,\theta-\rho}(t) \cos(l_{\alpha,\theta-\rho}(t))] \mathrm{d}t,$$

where $k_{\alpha,\phi}(t)$ and $l_{\alpha,\phi}(t)$ are defined in (11). Note that the function $a(\rho)$ is defined for $\rho \in]-\infty, \theta[$. Moreover, let

$$a_1 = a(-\rho_1), \qquad a_2 = a(\rho_2), \qquad v = a(0), \tag{16}$$

$$b_1 = \min\left(\frac{1}{\rho_1+\rho_2}\log\frac{a_2}{a_1}, \frac{1}{\rho_1}\log\frac{v}{a_1}\right), \; b_2 = \max\left(\frac{1}{\rho_1+\rho_2}\log\frac{a_2}{a_1}, -\frac{1}{\rho_2}\log\frac{v}{a_2}\right), \tag{17}$$

$$w_1 = \frac{a_1\exp(\rho_1 b_1)}{A\rho_1}, \; w_2 = \frac{a_2\exp(-\rho_2 b_2)}{A\rho_2}, \; w_3 = \frac{v(b_2-b_1)}{A}, \tag{18}$$

where

$$A = A(\rho_1, \rho_2) = \frac{a_1\exp(\rho_1 b_1)}{\rho_1} + \frac{a_2\exp(-\rho_2 b_2)}{\rho_2} + v(b_2 - b_1) \tag{19}$$

is the rejection constant of the algorithm, i.e. the expected number of iterations in the algorithm. We choose ρ_1 and ρ_2 as $(\rho_1^*, \rho_2^*) = \arg\min A(\rho_1, \rho_2)$, where minimization is taken under the constraint on $a(\rho)$.

Our proposed algorithm for simulating the Tweedie distribution as an exponentially tilted stable law, which makes use of the quantities defined in Equations (16)–(19), is detailed in Appendix A as Algorithm 1. It is apparent that our algorithm is efficient in terms of the rejection constant $A(\rho_1^*, \rho_2^*)$, which is computed in Table 2 for selected values of α, θ and δ. It is seen that the rejection constant of the algorithm suggested by Devroye (2009) is always larger (and often much more so) than that for Algorithm 1, in spite of the fact that we have provided our method in a very basic form. In fact, our algorithm could be further improved by using clever computational tricks, such as "recycling" to save the generation of a uniform random variate in the loop, and suitable "squeezes" based on the characteristic function of X to skip the direct computation of f_X most of the times. In any case, direct evaluation of f_X may be eventually avoided by means of the series method (Devroye, 1986), even if at the cost of a more involved algorithm.

When $\alpha \in]-\infty, 0[$, the stochastic representation (3) yields the efficient algorithm given in Appendix A as our Algorithm 2. Indeed, the sum of i.i.d. copies of Gamma r.v.'s is in turn Gamma distributed, so that Algorithm 2 only requires generation of Poisson and Gamma variates, which are widely available.

6. Applications to international trade

This section describes two relevant sets of products on which we have satisfactorily fit the Tweedie model. Being able to anticipate the distribution of traded quantities for these products is important both for policy support and anti-fraud purposes. Other distributions have also been considered for the same data and the comparison is discussed in §7, together with additional insights on the fit, residual analysis and some computational details. The Matlab code that was used to analyze the data is available from the authors on request.

6.1. Petroleum oils

International political economy is especially interested in the 27th chapter of the CN, including "Mineral fuels, mineral oils and products of their distillation; bituminous substances; mineral waxes". In the period considered the chapter alone covered more than 27% of all EU imports. Within the chapter, we select the product of largest share (around 17% of all EU trade imports), i.e. "Petroleum oils and oils obtained from bituminous minerals,

crude" (CN-27090090). The EU is highly import-dependent for this product, with major impact on the balance of payments. There is also academic interest for this product. Crude oil cannot be studied under the classical price theory assumptions relying on perfect competition. In fact, the product is quite heterogeneous, with an actor, OPEC, with big market power but not behaving as a typical cartel (the OPEC members often pursue different policy goals). The volume of available reserves is unclear and oil demand is variable, even within the EU. Furthermore, oil prices are greatly influenced by the rapid growth of demand from emerging countries and by the availability of unconventional resources (shale oil). There is little agreement in economic research about how all these factors impact on the demand and, therefore, on the oil quantities imported in the EU. In this complex context, sound empirical approaches are clearly needed to support economic modelling of demand and supply. Our trade quantity model offers a tool for assessing and validating different price-quantity proposals.

Descriptive statistics of the Petroleum oil quantities, reported in Appendix B, give an idea of the order of magnitude of this trade. Perhaps surprisingly, given the nature of the product, a non-negligible amount of quantities are recorded as zero, in order to ensure confidentiality of the traders involved in such transactions. However, these null values cannot be modelled from a purely economic perspective since the precise incidence of confidentiality-related underreporting is unknown and regulations vary considerably from one market to another. The null quantities, together with the spike of transactions of relatively small quantities, are likely to influence the fit of any distribution to such data. Figure 2 displays the fit of the Tweedie model and the estimated parameters. Visual inspection shows the fit to be satisfactorily good. Indeed, the Tweedie distribution is effectively able to capture both the long tail of the empirical distribution and the considerable mass of observations close to zero. Furthermore, the negative value of $\hat{\alpha}$ is coherent with the qualitative interpretation of the nature of the jumps that are expected to give rise to the traded quantities of this product. Further quantitative insight on the fit is reported in §7.

6.2. Fraud sensitive products

Many different categories of goods were seized in the last years by OLAF investigations and joint customs operations. We select a number of particularly sensitive products, from numerous chapters of the CN, among those that have been object of official EU press releases, which can be retrieved

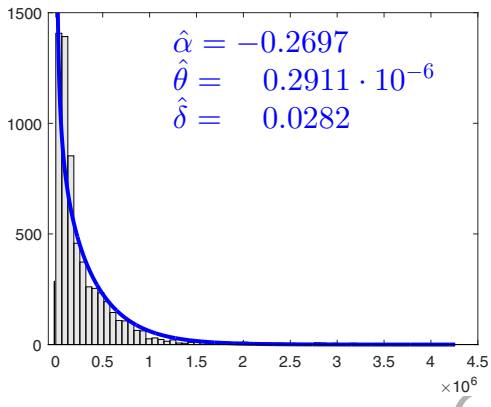


Figure 2: Fit of the Tweedie model to the traded quantities (in million of tons) of product "Petroleum oils and oils obtained from bituminous minerals, crude" (CN-27090090), already shown in Figure 1.

from the database `http://europa.eu/rapid` of the European Commission. We also classify the products according to the main fraud area to which they were associated by the press releases, excluding those at risk of smuggling (e.g., cigarettes and garlic), because the fraudulent amounts of such products escape from the data recording process. Our first category is very heterogeneous, and is related to the *counterfeiting* problem. It includes products such as wine (in CN-2204), alcohol (in CN-2208), textiles (from CN-50 to CN-63), footwear (in CN-64), electrical and electronic apparatus (in CN-84 and CN-85), accessories of motor-vehicles (mostly in CN-8708) and watches (from CN-9108 to CN-9110). A second less heterogeneous category, with more than 80 fruit and vegetable products in CN-20, relates to the problem of *mis-declaration* of product or origin.

We discuss here the findings of our empirical analysis of four products, two for each fraud category, but the results are similar for most of the sensitive products that we have classified. In order to study also the effect of the zeros on the fitted model, for each fraud category we select one product for which many null quantities are recorded and one without zeros. *Combed wools* and *Wines* with Protected Designation of Origin (PDO) or Protected Geographical Indication (PGI) are major products at risk of counterfeiting. *Mushrooms* are often improperly classified for evading duties, or their origin is misdeclared for bypassing the authorised exporter's quotas. In particular, many false classifications were reported for the 'Agaricus' type, even with

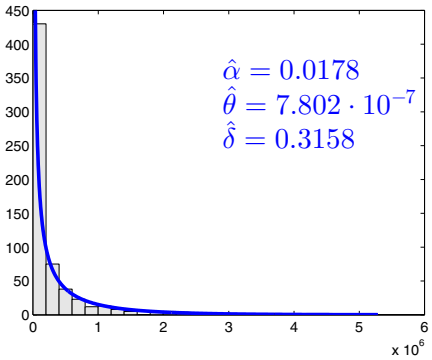
implications on food safety. *Cherries* are an example of fruit found to contain or being preserved with sugar or alcohol, but declared otherwise. This relates to duty evasion but also to illicit sugar trade practices.

Descriptive statistics for the four selected fraud sensitive products are in Appendix B, while Figure 3 shows the fit and the estimated parameters. From the histograms it is clear that the products have different data distributions. In two cases – depicted in plots (b) and (d) – there is also a non negligible mass of exactly null quantities that determine, as in the application to Petroleum oils, a negative estimate of parameter α. In the two other instances – plots (a) and (c) – all quantities are strictly positive, but there is a very large amount of observations close to the origin, thus implying that rounding to zero and confidentiality issues have a negligible impact for these products. As a result, we obtain positive estimates of α, which are coherent with the idea of a large number of almost instantaneous transactions typical of consumer goods, while the estimated values of θ are very close to zero. This suggests a stable-type shape of the empirical distribution which is automatically captured by our Tweedie model, without any tuning intervention. We thus see that, independently of the specific data generating mechanism, the Tweedie distribution is able to fit effectively both sides of the empirical distribution of traded quantities in all the selected anti-fraud examples.

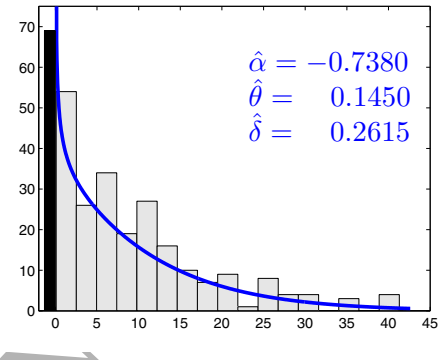
7. Comparison and further insights on the applications

We now compare the fit of the Tweedie and that of other five candidate models. We define the alternative models by mixturing some popular skew distribution functions for positive values, listed in Table 3, with a Dirac mass at the origin. In most cases the skew distribution functions that we select were successfully applied to describe economic patterns, such as income distribution (see, e.g., Clementi and Gallegati, 2005; Fisk, 1961). The Gamma distribution is one of such alternative candidates. As described in §4, it is also a special case of the Tweedie model and, for that reason, it has been considered as a potential competitor in several application fields (see, e.g., Palmer et al., 2008).

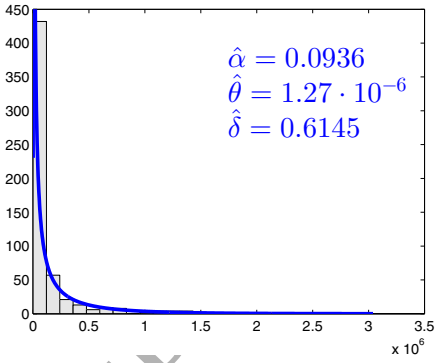
Classical models for skew distributions cannot be directly applied to products for which null quantities are observed. Since our goal is to provide comparison among models that could be routinely applied to international trade data without tuning intervention, we augment the standard two-parameter models with a spike at the origin. We thus fit the three-parameter mixture



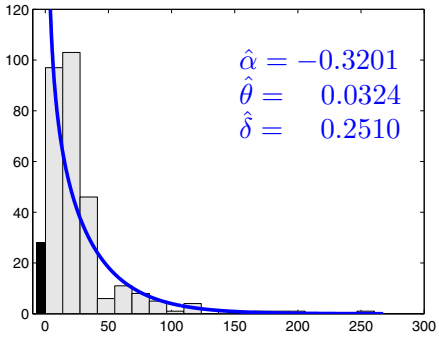
(a) Wine – CN-22042994



(b) Wool – CN-51052100



(c) Mushrooms – CN-20031030



(d) Cherries – CN-20086090

Figure 3: Fit of the Tweedie model to the traded quantities of fraud sensitive products. Product (a) is measured in liters, (c) in kilogram drained net weight and the other two products in tons. Data for products (a) and (c) do not contain null quantities. Products (b) and (d) contain, respectively, 23% and 8.9% of nulls (black bin of each histogram). The fitted parameters are highlighted in each plot.

distribution

$$F_X(x) = \pi I_{[0,\infty[}(x) + (1-\pi)G_X(x), \tag{20}$$

where G_X is the distribution function of any of the two-parameter random variables listed in Table 3 and $\pi \in]0,1[$ is an additional parameter which represents the probability of observing a quantity rounded to zero. Therefore, this mixture may be seen as a somewhat ad hoc three-parameter extension of traditional models for skew distributions, defined in view of the specific structure of trade data and mimicking the stochastic representation (3) of the Tweedie distribution.

We stress that it would be unfair to compare our approach with the standard two-parameter versions of the distributions given in Table 3, since they would not be able to allow for the presence of null values in many data sets. On the other hand, we frame our comparison in a "worst-case" scenario for the Tweedie model, since mixture (20) exploits some prior information about the structure of the data which is not available for the Tweedie distribution, placing the Dirac mass exactly where it is required, i.e. at the origin. Our reference distribution is also more parsimonious than the four-parameter generalized skew normal distribution proposed by Mazzuco and Scarpa (2015) to allow for bimodality in asymmetric patterns, which cannot model in any case the probability mass at zero. Incidentally, it should be remarked that the Tweedie distribution may even display bimodality, in addition to the automatic modelling of the zero mass for $\alpha \in]-\infty, 0]$ (see Aalen, 1992). The Tweedie distribution itself could be used as an ingredient in mixture (20), if a less parsimonious four-parameter model is required due to the complexity of data. However, we do not investigate this possibility in detail since our goal is to compare relatively simple three-parameter alternatives for routine analysis of trade data sets.

We estimate the three parameters in mixture (20) by means of Maximum Likelihood. We compare all models in terms of their maximized loglikelihood values and well-known divergence measures in Table 3, where the best values for each criterion are highlighted in bold. This allows an exhaustive analysis of the accuracy of the fit. In fact the likelihood evaluates the quality of the fit of the density function, while the other measures evaluate the divergence in terms of the distribution function. In particular, we use the Kolmogorov-Smirnov distance, which is perhaps the most natural candidate in the present context, and the Anderson-Darling distance, which places more weight on observations in the tails of the distribution. Since the Tweedie distribution

function cannot be written in closed form, it is computed from samples of 10,000 randomly generated observations, simulated using the random number generator described in §5.2 and the parameter estimates obtained for each product. Note that the optimal model in terms of Maximum Likelihood also optimizes the BIC and AIC comparison criteria, given that all the alternative models considered in our comparison have the same number of parameters and that the sample size is fixed for each product. Moreover, since mixture (20) is not nested in the Tweedie model, the likelihood ratio tests of Palmer et al. (2008) cannot be carried out to formally compare the Tweedie model to its competitors.

In terms of maximized loglikelihood, our results show that the Tweedie distribution provides the best fit, and in some cases by a large extent, while there is no clear evidence on the ranking of alternative models. For Combed Wool and Preserved Cherries, the fit of the Gamma version of the three-parameter mixture (20) is very close to that of the Tweedie distribution. However, in other cases it is the Weibull distribution which comes second, thus showing that it is not straightforward to specify a flexible and simple alternative to the Tweedie model which holds for all these data. Divergence measures confirm the general good fit of the Tweedie model. It is undoubtedly the best model for Combed Wool and Preserved Cherries. It can be considered the best performer also for Mushrooms Agaricus, since it optimizes the Anderson-Darling distance, which places more importance on the tail of the distribution, and it is very close to the best values for the other two divergences. For the remaining two products, the performance of the Tweedie fit is very close to the correspondent best model.

As a further drawback of mixture (20), we note that for two products (Wine and Mushrooms) and the Lognormal version of mixture (20) the Matlab estimation algorithm failed to converge, despite our careful tuning of tolerance options. We take this outcome as a further drawback of standard Maximum Likelihood when applied to mixture (20), where there is a significant mass probability at one point in the domain of the distribution and the method may become unstable.

On the other hand, the Maximum Likelihood approach to the Tweedie model described in §5.1 did not suffer from computational problems. Specifically, we truncated the series involved in (5) at 170 terms, although a much smaller number – such as 20 or 30 – was generally sufficient for achieving convergence in most of the simulations we performed and data sets we analyzed. The Matlab function "mle" with options 'MaxIter'=1000 and 'MaxFu-

Table 3: Comparison of maximized loglikelihood values and divergence measures for the fit on the Petroleum oils and fraud-sensitive data sets, based on different skew distributions G_X in (20) and on the Tweedie distribution (LL = Loglikelihood, AD = Anderson-Darling distance, KS = Kolmogorov-Smirnov distance).

		Distribution G_X in (20)					
		Gamma	**Lognorm.**	**Inverse Gamma**	**Loglog.**	**Weibull**	**Tweedie**
Wine PDO	**LL**	-7,936.7	-	-8,329.7	-8,047.6	-7,959.1	**-7,930.7**
or PGI	**AD**	**4.145**	-	11.299	24.419	91.791	4.956
(22042994)	**KS**	**0.081**	-	0.110	0.127	0.331	0.088
Combed	**LL**	-900.7	-938.4	-1,024.8	-929.8	-901.6	**-897.6**
wool	**AD**	10.361	14.233	10.226	12.005	26.863	**9.572**
(51052100)	**KS**	0.234	0.234	0.234	0.234	0.234	**0.229**
Mushrooms	**LL**	-6,638.8	-	-6,921.3	-6,674.8	-6,627.6	**-6,621.0**
agaricus	**AD**	3.644	-	1.262	7.594	70.920	**1.090**
(20031030)	**KS**	0.071	-	**0.043**	0.071	0.279	0.054
Preserved	**LL**	-1,255.6	-1,301.5	-1,381.8	-1,305.0	-1,263.4	**-1,255.1**
cherries	**AD**	8.599	17.380	10.446	14.776	29.870	**6.924**
(20086090)	**KS**	0.196	0.258	0.210	0.213	0.307	**0.165**
Petroleum	**LL**	-87,545.3	-89,454.6	-97,597.6	-88,003.9	-87,539.8	**-87,526.1**
oils	**AD**	58.961	292.709	**46.650**	51.367	1,905.606	75.726
(27090090)	**KS**	0.072	0.164	0.074	**0.064**	0.463	0.078

nEvals'=2000 was used to maximize the likelihood. The starting values of the parameters were selected according to the moments of the empirical distribution, and eventually perturbed in case of no convergence of the algorithm. With these settings we did not experience convergence problems in the fitting process of the Tweedie distribution. Furthermore, we double-checked our results by implementing the likelihood equations as Mathematica functions. Equations (12) and (15) obviously involve integrals whose integrands are oscillating functions, in such a way that their oscillating behaviour is ruled by the parameter ϕ. As suggested in §5.1, this parameter was generally set to unity and the required integrals were computed by means of the numerical integration routines provided by Mathematica in the standard setting (i.e., without adopting special tunings or options for oscillating integrands). The achieved values of the Maximum Likelihood estimates consistently matched for the two softwares, also in the case of quite extreme data sets with a considerable mass at zero. Therefore, we argue that the Tweedie model appears to be more stable than mixture (20) also from a computational point of view.

We supplement our comparisons through some residual analysis on the

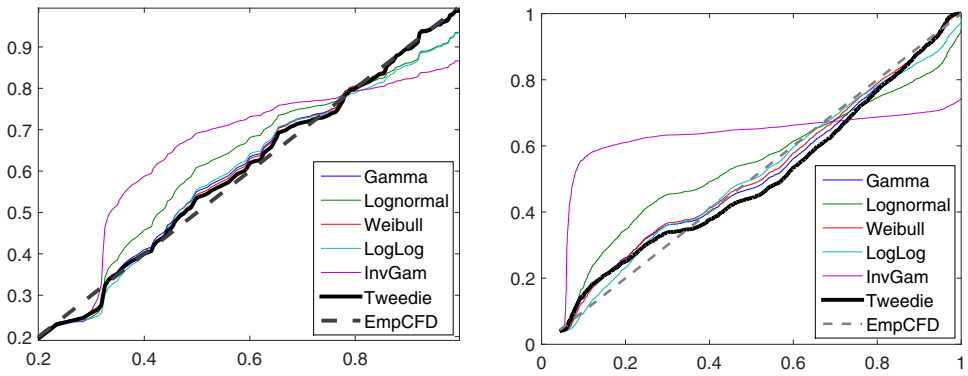


Figure 4: *QQ*-plots for the six fitted models against the quantiles of the empirical distribution function (the dashed straight line on the diagonal) for two products: Combed wool (left panel) and Petroleum oils (right panel).

fit of the different models. Figure 4 reports the estimated quantiles under the six fitted models against the quantiles of the empirical distribution function (reported as the dashed straight line on the diagonal of each plot) for two products: Combed wool and Petroleum oils. Due to the spike at the origin, we do not plot the quantiles associated to probabilities smaller than 0.2. Furthermore, to facilitate interpretation, the estimated quantiles of the Tweedie model are represented with a thick black solid curve. Again, the Tweedie quantiles are computed from samples of 10,000 randomly generated observations, simulated using the random number generator described in §5.2 and the parameter estimates obtained for each product.

We cannot expect one single model to uniformly dominate all the others over the entire support of the distribution. Nevertheless, it is seen that the fit of the Tweedie distribution is generally good and it is especially so in the right tail of the empirical distribution, even in the case of Petroleum oils for which other candidate mixtures (20) exhibit slightly lower divergences. These *QQ*-plots thus provide visual confirmation that the excellent global performance of the Tweedie model in in Table 3 is not marked by isolated deviations due to some specific transactions.

From our comparisons, we conclude that the Tweedie distribution offers an effective modelling solution even when compared to mixtures that incorporate prior information on the structure of the data, and that may be thus favoured in the analysis of traded quantities. Its very flexible parametrization

allows a simple and statistically sound representation of empirical data which can be useful in many anti-fraud and policy support settings. It also qualifies the Tweedie distribution as an ideal candidate for Monte Carlo simulations mimicking the real structure of international trade data.

8. Stability over time

The Tweedie parameters estimated in §6 for Petroleum oils and the other fraud-sensitive products are based on data covering a four-years period (i.e. August 2008 - July 2012). This time-length interval is usually classified in economics as "middle-term". In order to validate the good behaviour and the proper fitting of the Tweedie model also in the short-term, we have repeated the analysis on five shorter and consecutive sub-periods, at least for the products that guarantee a sufficient number of observations in each sub-period (namely Petroleum oils, Wine PDO and PGI protected and Mushrooms Agaricus). We provide here only the case of Petroleum oils (Table 4), since results and conclusions for the other products are very similar.

The sub-period analysis highlights two important facts. First, the Tweedie model still offers the best fitting in terms of maximized loglikelihood (and then also in terms of BIC and AIC) in most of the sub-periods considered, when compared to the alternative candidate mixtures (20). Moreover, it guarantees also the highest average loglikelihood per observation, a value that can be considered for summarizing the general performance of a model over the five sub-periods. Second, the parameter estimates remain stable and do not appreciably differ from the values estimated for the whole period (see Figure 2). This limited volatility of the coefficients reflects the substantial stability of the quantity distribution over the five sub-periods, as the minor changes in the descriptive statistics proves. Furthermore, following the discussion at the end of §4.4, it confirms that the trade mechanisms operating in these markets are essentially stable over time. So we can conclude that the choice of a specific time window does not affect the main substantial findings, as is desirable in our operative context.

9. Conclusions

We have shown the wide applicability to international trade data of a flexible parametric model based on the Tweedie distribution. Our proposal

Table 4: Annual estimates and comparison of maximized loglikelihood values for the fit on "Petroleum oils and oils obtained from bituminous minerals, crude".

	2008[a]	**2009**	**2010**	**2011**	**2012**[a]	
			Descriptive statistics			
# of obs.	717	1618	1667	1702	947	
# of zeros	12	29	66	109	70	
mean	318,971.3	313,040.8	305,010.4	282,145.3	296,727.4	
std dev.	469,937.2	472,347.1	485,961.8	436,475.4	448,494.1	
skewness	3.44	3.68	3.79	3.83	3.58	
kurtosis	17.78	19.74	20.76	22.87	19.63	
			Tweedie parameters estimates			
$\hat{\alpha}$	-0.1974	-0.2155	-0.2583	-0.2795	-0.2820	
$\hat{\theta}$	$2.68 \cdot 10^{-6}$	$2.82 \cdot 10^{-6}$	$2.64 \cdot 10^{-6}$	$2.88 \cdot 10^{-6}$	$2.64 \cdot 10^{-6}$	
$\hat{\delta}$	0.0679	0.0606	0.0317	0.0229	0.0209	
			Annual comparison			**Mean Loglik. per obs.**[b]
Gamma	-9,643.5	-21,800.5	-22,024.4	-21,904.8	-12,113.8	-13.159
Lognormale	-9,836.2	-22,044.1	-22,412.5	-22,484.2	-12,482.4	-13.436
Weibull	-9,643.9	-21,781.0	-22,010.4	-21,919.2	-12,133.1	-13.161
Loglogistic	-9,699.5	-21,815.3	-22,088.8	-22,079.0	-12,246.1	-13.233
Inv.Gamma	-10,741.8	-24,103.2	-24,532.1	-24,392.8	-13,481.5	-14.633
Tweedie	-9,641.5	-21,835.8	-22,042.8	-21,876.5	-12,083.0	-13.156

[a] For 2008, data cover the period August 2008 - December 2008, whereas for 2012 data refer to the period January 2012 - July 2012.
[b] This value is given by the mean over the five sub-periods of the ratios between the value of the loglikelihood and the number of observations.

is particularly suitable for products of paramount importance in the EU market, especially for anti-fraud purposes. To this aim, we have solved a number of tricky statistical issues concerning model parametrization, estimation and random variate generation. These issues are important both for practical routine implementation of our model and for performing Monte Carlo evaluation of the properties of anti-fraud tools.

We have developed this work in support of the EU decision-makers and law enforcement services. Nevertheless, we believe that the applicability of our model is much wider. For example, the precise description of the trade dynamics that it provides can help both national services in estimating trade balance and private companies in simulating the effect of market dynamics. Our focus has been on modelling the trade quantity, but this variable has direct, often linear, relation with capital flows and prices. Economic theory has

widely studied how sensible fluctuations in the exchange rates can determine reactions in the trade prices and, in turn, changes in the trade quantities (see, e.g., Goldberg and Knetter, 1997; Arkolakis et al., 2012; Burstein and Gopinath, 2015). Economic research has addressed these relationships with particular attention to strategic trade areas affected by imperfect competition, such as the petroleum oils. To be concretely applicable to trade, monetary and inflation policies, these studies need to be supported by empirical evidence that we argue can be found in data and models such as those addressed in this paper.

In addition to the stated goals of policy support and data simulation, the availability of a close-to-reality model for trade quantities can also help to investigate how different conditions, or structural changes, in the international economy affect transactions in the EU market. As we have shown, this goal can be achieved through seeing whether the estimated values of the parameters of the Tweedie distribution remain stable over time, especially when different economic phases are considered. Similarly, the estimated parameters of the Tweedie model could be used to check homogeneity of trade within the same CN chapter.

We close this discussion of other potential applications of our model with a mention of the COMTRADE database (http://comtrade.un.org), also formed by monthly aggregates built on the basis of the UN protocols. COMTRADE is geographically broader than COMEXT, as it collects data from all members of the United Nations, but the products are defined only at 6-digits level in the Harmonised System (HS) classification. Therefore, COMTRADE contains less sub-headings than COMEXT (about 6000). It will be an interesting task for future research to investigate whether the approach presented in this work is applicable to these "less precise" aggregates or, in other words, to check at what extent the granularity of the classification impacts on the estimated model.

Acknowledgements

This work was jointly supported by the Project "Automated Monitoring Tool on External Trade, Step4" of the Joint Research Centre and the European Anti-Fraud Office of the European Commission, and by the project MIUR PRIN "*MISURA – Multivariate models for risk assessment*". We are grateful to Anthony Atkinson, Daniela Buscaglia, Mario Menegatti and Marco Riani for helpful discussion on previous drafts of this work. Finally,

we thank the Editor and two anonymous reviewers for several helpful comments and for suggesting additional analyses that have enhanced the scope of the work.

Appendix A: Simulation Algorithms

Algorithm 1: Tweedie distribution with $\alpha \in]0,1[$

```
compute v, ρ1*, ρ2*
compute a1, a2, b1, b2
compute w1, w2, w3
repeat
    generate U1, U2, U3 uniformly on ]0,1[
    if U1 > w1 + w2 set X := b1 + (b2 − b1)U2
    else
        if U1 ≤ w1 set X := log U2/ρ1* + b1
        else
            set X := − log U2/ρ2* + b2
until fX(X) < min {a1 exp(ρ1*X), a2 exp(−ρ2*X), v} U3
return X
```

Algorithm 2: Tweedie distribution with $\alpha \in]-\infty,0[$

```
input α,θ,δ
set X = 0
generate N Poisson with parameter (−δθ^α/α)
if N > 0 generate X Gamma with shape parameter (−Nα) and scale parameter
   (1/θ)
return X
```

Appendix B: Descriptive statistics for 'Petroleum oils' and other four fraud-sensitive products in the whole time window August 2008 – July 2012

The statistics in the table below are computed on imports in the EU in the period from August 2008 to July 2012. Petroleum oils, Combed wool and Preserved cherries are measured in tons, Wine in liters and Mushrooms in kilogram drained net weight. The first two fraud-sensitive products are at risk of counterfeiting while the last two are at risk of origin or product mis-declaration.

Description	CN8 code	Number of obs.	Number of zeros	mean	std dev.	skewness coeff.	kurtosis coeff.
Petroleum oils	27090090	6651	286	301440	463360	3.7145	20.5782
Wine PDO or PGI protected	22042994	633	0	323165.65	694815.51	3.98	21.34
Combed wool in fragments	51052100	295	69	7.50	8.93	1.53	5.12
Mushrooms agaricus	20031030	567	0	135883.13	321331.56	4.35	26.59
Preserved cherries	20086090	314	28	23.19	32.17	3.24	18.07

References

Aalen, O. (1992). Modelling heterogeneity in survival analysis by the compound poisson distribution. *Annals of Applied Probability 2*, 951–972.

Arkolakis, C., A. Costinot, and A. Rodriguez-Clare (2012). New trade models, same old gains? *American Economic Review 102*, 94–130.

Babaei, S., M. Sepehri, and E. Babaei (2015). Multi-objective portfolio optimization considering the dependence structure of asset returns. *European Journal of Operational Research 244*, 525–539.

Barabesi, L. and L. Pratelli (2014). A note on a universal random variate generator for integer-valued random variables. *Statistics and Computing 24*, 589–596.

Barabesi, L. and L. Pratelli (2015). Universal methods for generating random variables with a given characteristic function. *Journal of Statistical Computation and Simulation 85*, 1679–1691.

Barndorff-Nielsen, O. and N. Shephard (2001). Normal modified stable processes. *Theory of Probability and Mathematical Statistics 65*, 1–19.

Bianchi, M. and F. Fabozzi (2014). Discussion of 'On simulation and properties of the stable law' by Devroye and James. *Statistical Methods and Applications 23*, 353–357.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability 31*, 929–953.

Burstein, A. and G. Gopinath (2015). International prices and exchange rates. In E. Helpman, K. Rogoff, and G. Gopinath (Eds.), *Handbook of International Economics. Volume 4*, pp. 391–451. Elsevier.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association 105*, 147–156.

Cerioli, A. and D. Perrotta (2014). Robust clustering around regression lines with high density regions. *Advances in Data Analysis and Classification 8*, 5–26.

Clementi, F. and M. Gallegati (2005). Pareto's law of income distribution: evidence for Germany, the United Kingdom, the United States. In A. Chatterjee, S. Yarlagadda, and B. Chakrabarti (Eds.), *Econophysics of wealth distributions*, pp. 3–14. Milan: Springer.

Council (1995). Regulation (EC, Euratom) No 2988/95 of 18 December 1995 on the protection of the European Communities financial interest. EUR-Lex, http://eur-lex.europa.eu. Official Journal of the European Union, L 312, 23.12.1995, p. 1-4.

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.

Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation 19*, Article 18.

Devroye, L. and L. James (2014). On simulation and properties of the stable law. *Statistical Methods and Applications 23*, 307–343.

Dunn, P. and G. Smyth (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing 15*, 267–280.

Dunn, P. and G. Smyth (2008). Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Statistics and Computing 18*, 73–86.

Erlander, S. (1977). Accessibility, entropy and the distribution and assignment of traffic. *Transportation Research 11*(3), 149 – 153.

Erlander, S. (1982). Accessibility, entropy and the distribution and assignment of traffic revisited. *Transportation Research Part B: Methodological 16*(6), 471 – 472.

European Commission (2012). Communication to the European Parliament, the Council and the European Economic and Social Committee on Customs Risk Management and Security of the Supply Chain. EUR-Lex, http://eur-lex.europa.eu. COM(2012) 793.

Eurostat (2006). Statistics on the trading of goods - user guide. URL: http://epp.eurostat.ec.europa.eu/. ISSN 1725-0153, ISBN 92-79-01577-X.

Favaro, S. and B. Nipoti (2014). Discussion of 'On simulation and properties of the stable law' by L. Devroye and L. James. *Statistical Methods and Applications 23*, 365–369.

Fisk, P. (1961). The graduation of income distributions. *Econometrica 29*, 171–185.

Fogelman-Soulie, F., D. Perrotta, J. Piskorski, and R. Steinberger (2008). *Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and Their Applications to Security*. Amsterdam, The Netherlands, The Netherlands: IOS Press.

Francq, C. and J.-M. Zakoian (2013). Estimating the marginal law of a time series with applications to heavy-tailed distributions. *Journal of Business and Economic Statistics 31*, 412–424.

Fray, H. and A. G. Wilson (2012). A dynamic global trade model with four sectors: food, natural resources, manufactured goods and labour. Working Paper Series 178, University College London, Centre for Advanced Spatial Analysis, London.

Goldberg, P. and M. Knetter (1997). Goods prices and exchange rates: What have we learned? *Journal of Economic Literature 35*, 1243–1272.

Harris, G. (1974). *Regional Economic Effects of Alternative Highway Systems*. Cambridge, Mass.: Ballinger Publishing Co.

Hofert, M. (2011). Sampling exponentially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation 22*, Article 3.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika 73*, 387–396.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics 4*, 227–241.

Jörgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society, Series B 49*, 127–162.

Kapur, J. N. (1993). *Maximum-Entropy Models in Science and Engineering (Revised Edition)*. John Wiley & Sons. First Published in 1989.

Lejeune, I., R. Tusveld, D. Aerts, M. Wagemans, N. Bogaerts, and C. Buysing Damste (2013). Study on the evaluation of the customs union. Technical report, DG TAXUD. ISBN 978-92-79-33136-7, DOI 10.2778/17430.

Lijoi, A. and I. Prunster (2014). Discussion of 'On simulation and properties of the stable law' by L. Devroye and L. James. *Statistical Methods and Applications 23*, 371–377.

Mazzuco, S. and B. Scarpa (2015). Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. *Journal of the Royal Statistical Society, Series A 1*, 107–203.

Menn, C. and S. Rachev (2005). A GARCH option pricing model with -stable innovations. *European Journal of Operational Research 163*, 201–209.

Palmer, K., M. Ridout, and B. Morgan (2008). Modelling cell generation times by using the tempered stable distribution. *Journal of the Royal Statistical Society, Series C 57*, 379–397.

Perrotta, D. and V. Kopustinskas (2010). Discussion: The forward search: Theory and data analysis. *Journal of the Korean Statistical Society 39*, 147 – 149.

Rachev, S., Y. Kim, B. M.L., and F. Fabozzi (2011). *Financial Models with Lévy Processes and Volatility Clustering*. Hoboken, New Jersey: Wiley.

Riani, M., A. C. Atkinson, and A. Cerioli (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B 71*, 447–466.

Ridout, M. (2009). Generating random numbers from a distribution specified by its Laplace transform. *Statistics and Computing 19*, 439–450.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto (2004). *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. Chichester: Wiley.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge, UK: Cambridge University Press.

TFEU (2012). Treaty on the functioning of the european union (consolidated version 2012). EUR-Lex, http://eur-lex.europa.eu/en/treaties/index.htm. Official Journal of the European Union, C 326, 26.10.2012.

Theil, H. (1967). *Economics and information theory*. Studies in mathematical and managerial economics. North-Holland Pub. Co.

Tweedie, M. (1984). An index which distinguishes between some important exponential families. In J. Ghosh and J. Roy (Eds.), *Statistics: Applications and New Directions, Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, pp. 579–604. Calcutta: Indian Statistical Institute.

Wilson, A. G. (1970). Inter-regional commodity flows: Entropy maximizing approaches. *Geographical Analysis 2*(3), 255–282.

Zhang, W.-G., Y.-J. Liu, and W.-J. Xu (2012). A possibilistic mean-semivariance-entropy model for multi-period portfolio selection with transaction costs. *European Journal of Operational Research 222*, 341–349.