Contributions for the development of the European Anti-Fraud Observatory: a statistical perspective

Maria Felice Arezzo

DIPARTIMENTO DI METODI E MODELLI
PER L'ECONOMIA IL TERRITORIO E LA FINANZA
MEMOTEF

CENTRO DI RICERCA
IMPRESAPIENS

SAPIENZA
UNIVERSITÀ DI ROMA

1st Workshop of the project "101101784 — 2022-IT-FRED2"
23 November 2023

Outline of the presentation

- Possible contributions of the Observatory to the national anti-fraud strategy (NAFS). A statistical perspective.
  - Estimation of size and determinants of frauds;
  - Risk profiles;
  - Planning of inspections;
  - Taxonomy of the irregularities;
  - Knowledge sharing;
  - Other?
- Methods
  - Statistics and machine learning algorithms
  - Issues in the data (inference for rare events, selection bias (?), misclassification(?))
- Structure of the Observatory

Estimation of the size of frauds

The estimation of the size of fraud can be seen from different angles:

- Estimation of the **proportion** of funded projects that are fraudulent;
- Estimation of the **number** of funded projects that are fraudulent;
- Estimation of the **proportion** of budget spent on fraudulent projects;
- Estimation of the **amount** of budget spent on fraudulent projects.

Each estimate (**point and interval**) shed lights on a specific aspect of the issue: for example, we can have a high number of frauds along with a low amount subtracted or a low proportion of fraudulent projects with an associated high amount of budget. These situations are very different.

Monitoring **significant** differences among programs, types of recipients, years, territory is useful for understanding and for policy recommendation.

Estimation of the determinants of frauds

A determinant in statistics is a variable that can be directly or indirectly measured and that is known or it is suspected to have a significant association with the variable of interest.

Complex phenomena, such as frauds, have several determinants.

The added value of the statistical analysis is that it allows to *quantify the magnitude and the direction* of the effect of *each* determinant on the variable of interest *all other things held constant*.

As such, it is possible to simulate how the variable of interest (i.e. the “fraud”) would change when a determinant change.

Risk profile

Closely related to the determinants is the risk profiling of each project.

A risk profile is a **quantification of the probability** that a project with certain characteristics is (or will be) a fraud.

Such an assessment is based on a process of learning from past experience.

Planning of the inspections

Inspections are one of the key elements in ensuring the proper use of EU funds.

In fact, a fraudster weights the benefits of fraud against the risks of being caught and fined (using the theoretical framework of Allingham and Sadmo, Income Tax Evasion: a Theoretical Analysis, 1972).

Inspection efficiency (i.e. planning inspections according to the risk of the project) ensures more effective monitoring for the same cost of inspection activities.

Beyond frauds: a taxonomy of the irregularities

What are the recurring characteristics of irregularities (if any)?

Would a taxonomy of irregularities make it possible to rank them in terms of severity?

Can we use such taxonomy to understand how to prevent their occurrence?

Other?

What else is of interest?

Methods. A brief and mostly harmless overview on how to: 1) estimate size

The most suited methodology depends on how data on frauds are collected, that is:

- The projects to audit are **chosen at random;**
- The projects to audit are **selected** based on their believed likelihood to be fraudulent;
- A mixed approach.

When we want to estimate the size of frauds (either in terms of proportion or amount) and the projects audited are a random sample, the estimation is straightforward (we use the corresponding sample statistics possibly weighted).

If the sample is not random, the sample statistics must be weighted so that the weights **take account of the selection process**.

2) Quantify the effect of the determinants

Once we have a list of possible determinants, we need to model the way they influence the variable of interest.

$$Y_i = f(\boldsymbol{X}_i) + noise$$

Y is the variable of interest (ex: the amount of fraud, the number of frauds, the occurrence of a fraud);

X is the ensemble of variables that are associated with Y;

f(·) is the way Y and **X** are linked together.

In the process of modelling data we must choose **ex ante** (colors reflects difficulty of the task):

- The variable of interest
- The determinants
- How to treat f(·): data driven or imposed by the researcher

The determinants: the fraud triangle

Why do people commit frauds?

There are several theories in psychology that provide an explanation on the mechanisms and motivations of fraudolent behavior.

A well known is the fraud triangle theory (Donald R. Cressey 1950's). The fraud triangle refers to the three elements that are typical precursors to fraudulent activity. The three elements, or legs, of the triangle include **opportunity**, **pressure** and **rationalization**

- **Opportunity:** is the joint action of the knowledge of ones position and the technical skills required to commit a fraud;
- **Pressure**: is a non-shareable financial pressure (for example falling behind of bills or wanting to have a higher-level lifestyle);
- **Rationalization**: how a potential fraudster justifies the crime before committing the fraud. Even when opportunity and financial pressure are present, many fraudsters feel the need to justify their actions in order to feel as though they are not social deviants

The determinants: the inestimable value of experience

Those who have worked on EU budget fraud for years have enormous experience and know what the determinants are.

The problem then is to **get hold** of such **measurable** determinants.

The enormous efforts done over the years have produced a remarkable data collection (Arachne):

With regard to improving the way data on detected frauds and irregularities are collected and used, 14 Member States reported they had fully implemented the related recommendation and had expanded their use of IT systems (Irregularity Management System (IMS), ARACHNE, and a further set of national IT tools).

(source: Relazione annuale COLAF Italia 2021)

The choice of f(·): researcher-driven approach

If we opt for a researcher-driven choice of f(·), then we can **quantify** how much a change in each determinant Xj impact the (expected) change in Y.

Example: suppose we want to model the Y=the amount of fraud (in euros) and we know that the determinants are X1, X2, X3. If we decide that f(·) is linear and each determinant has an additive effect the estimated model could be:

$$E(Y|X)=1500 + 2800X1 -3700X2 + 2500X3$$

So, experience tells us that **on average** X1 act as a risk enhancer and quantify the effect (2800 euros).

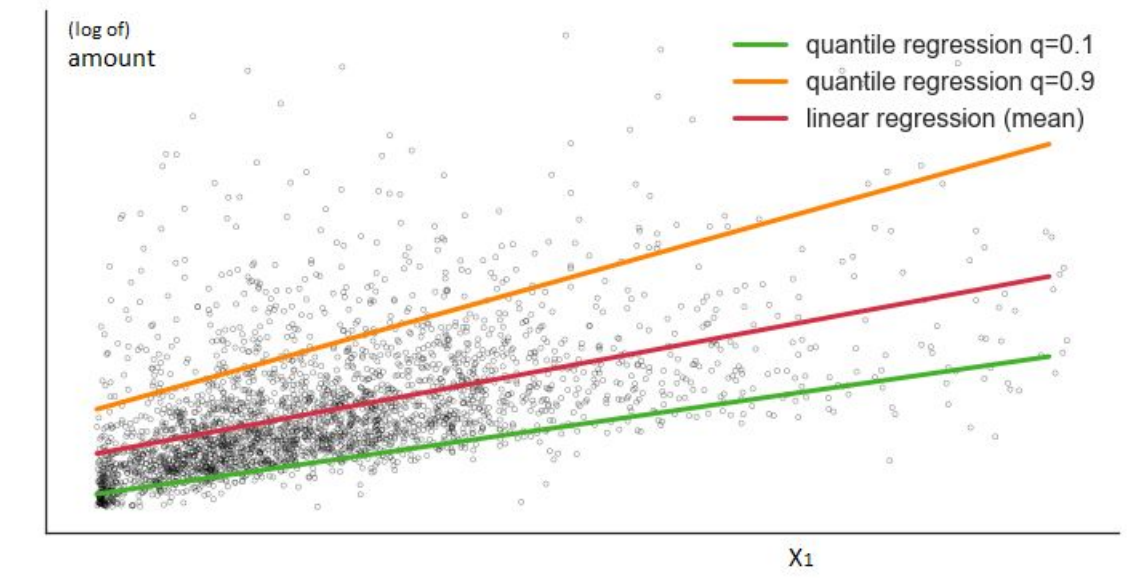
A similar interpretation applies –mutatis mutandis- to whatever variable of interest we are interested in.

Choosing f(·) is a **strong assumption** that is rewarded with a deeper understanding of the role of each determinants.

Statistical literature developed sophisticated models that can capture different nuances in the data.

One interesting example is the quantile regression that allows to understand the role of each determinant at different levels of the variable of interest.

The role of X_1 is different over the distribution of Y: an increase in X_1 has a higher impact in larger frauds (yellow line) than in modest ones (green line).



The choice of f(·): data-driven approach

In this approach, we do not get to quantify the impact of each determinant on the variable of interest. This loss of information is compensated by the avoidance of imposing possibly restrictive assumptions on data.

Input (**X**) → black box → output (prediction of Y)

Most known examples are:

- Random forests
- Neural networks
- Support vector machine

3) Derive risk profiles

Once we have a list of possible determinants, we need to model the way they influence the variable of interest.

$$Y_i = f(\boldsymbol{X}_i) + noise$$

Y is the variable of interest (ex: the amount of fraud, the number of frauds, the occurrence of a fraud);

X is the ensemble of variables that are associated with Y;

f(·) is the way Y and **X** are linked together.

In the process of modelling data we must choose **ex ante** (colors reflects difficulty of the task):

- The variable of interest
- The determinants
- How to treat f(·): data driven or imposed by the researcher

A word of warning…

Whenever we model data, we first need to study the characteristics of the data and be aware of any assumptions underlying the models we are using.

It does not exist anything such as “press a button and get a (reliable) result”.

In modeling frauds, we:

- Must keep in mind that we are dealing with **rare events**;
- Understand how data are collected (**random or non-random** inspections);
- Understand if all the variables at stake are **error-prone** or not;

4) Planning of inspections

Whether or not the inspection are random, statistical and mathematical thinking can be of great help.

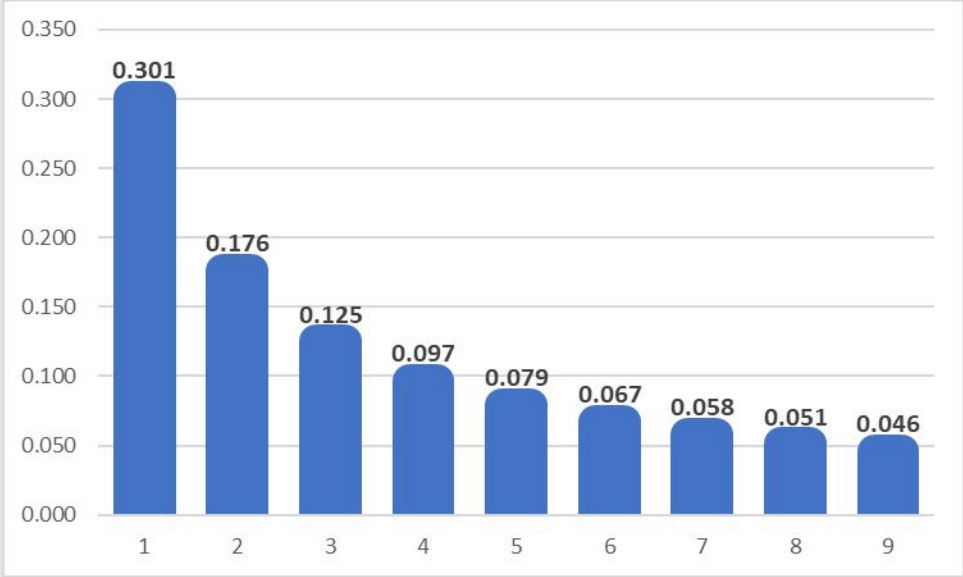
In the first case (random selection of projects to inspect), the sampling theory can design samples with the same level of reliability but lower number of projects to inspect.

In the second case (risk-driven selection of projects to inspect), along with the risk profiling that we discussed before, we can use the so called “empirical laws” such as the Benford’s law.

Benford’s law

The occurrence of digits 1 through 9 as the leftmost nonzero digit of numbers from real-world sources is distributed unevenly according to an empirical law, known as Benford's law or the first digit law.

Significant deviations from this distribution are possible symptoms of fraudulent behavior.



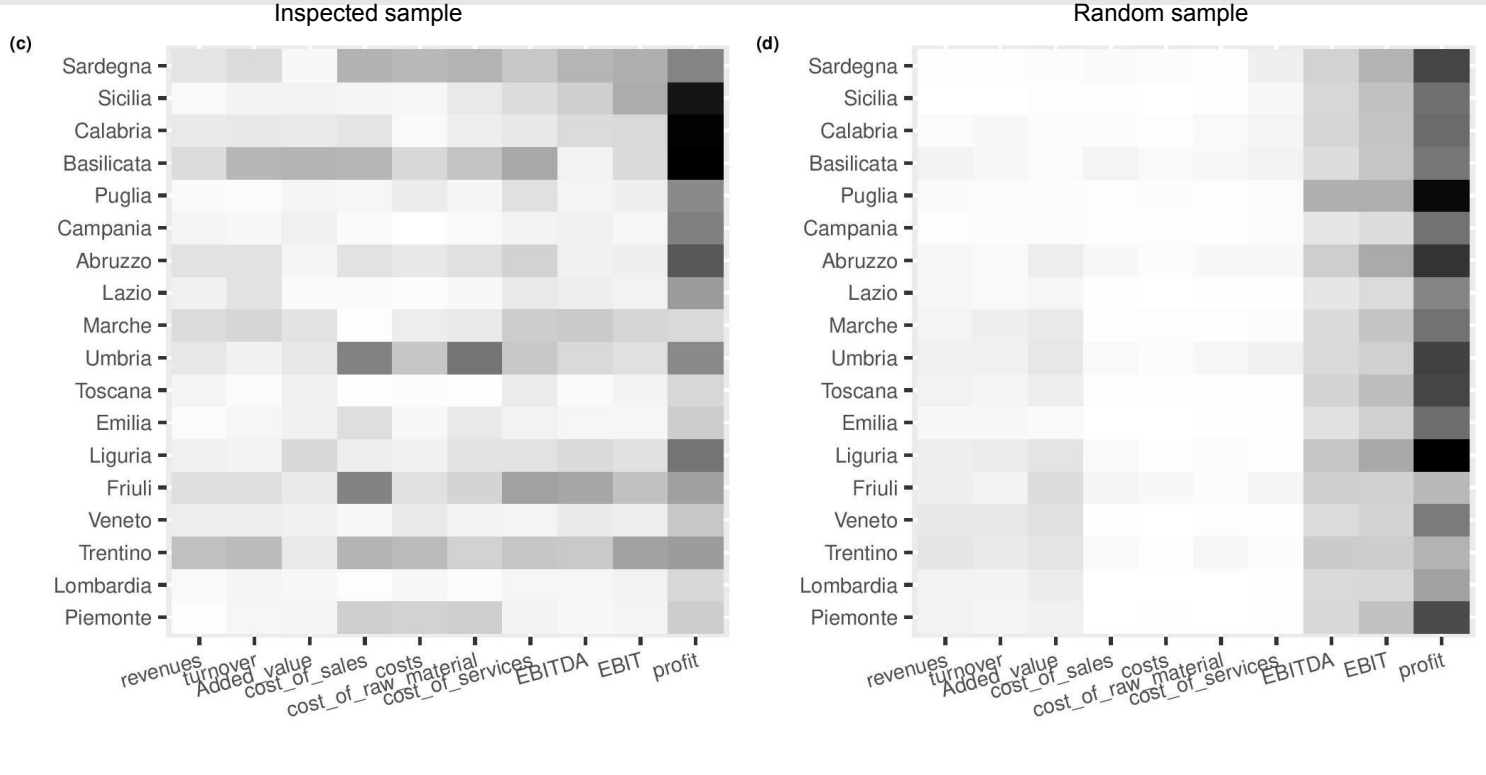
Benford law can be used, for example, to:

- Compare inspection activities in different territory/funds/years;
- Identify the budget items that more than other can suggest there is something to investigate

An example from a different field (Arezzo and Cerqueti, 2023, A Benford's Law view of inspections' reasonability):

The darker the less compliant to BL

By row (left panel): the darker the better
By column: look at the costs items in the balance sheet



5) Taxonomy of irregularities

The auditors find that the overall level of errors in spending from the EU budget increased in 2021, to 3.0 % (2020: 2.7 %). Nearly two thirds of the audited expenditure (63.2 %) was considered high-risk, also an increase compared to 2020 (59%) and before. The ***rules and eligibility criteria*** *governing this type of expenditure* ***are often complex****, which makes errors more likely. Material error continues to affect high-risk expenditure, at an estimated rate for 2021 of 4.7 % (2020: 4.0 %).* (Source: European Court of Auditors: annual report 2021)

In statistics, a taxonomy is created through a clustering process. Neither the number nor the composition of the clusters is known a priori, but is determined by studying the distances between the elements we want to cluster.

6) Knowledge sharing: outside-in and inside-out the observatory

- Frauds fight involves many players of the highest professional level, each with a technical language, knowledge, competences.

- Dissemination of the results:
  - Internal meetings;
  - Scientific conferences*;
- Research papers* to be published in academic journals and research reports;
- Creation of software for the implementation of analyses and statistical models.

*No results will ever be published if it violates privacy or confidentiality

Organization (draft) of the Observatory

Executive board (EB):

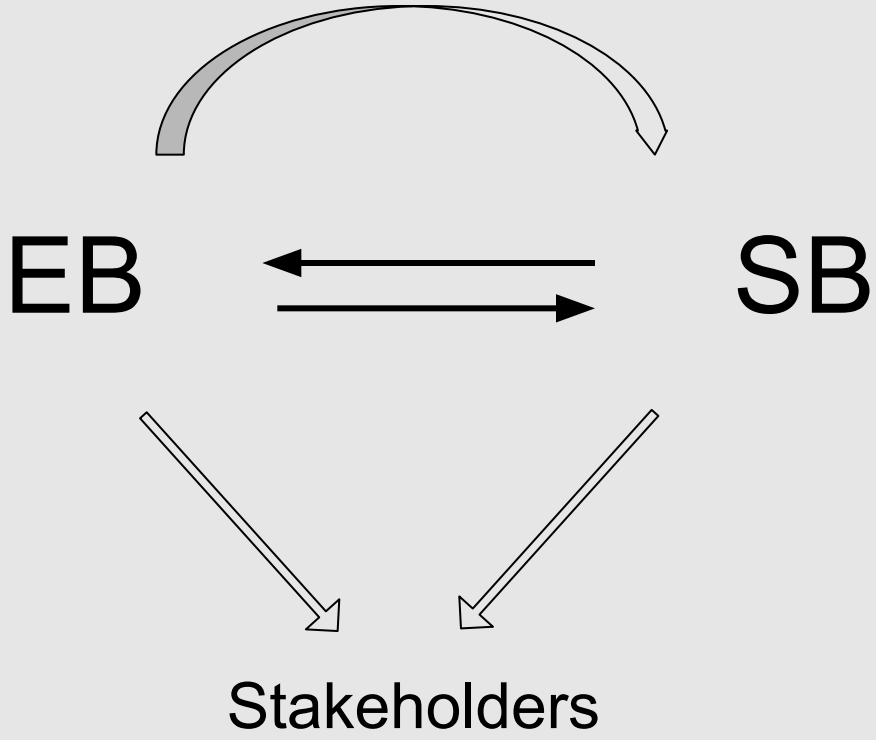
- 2/3 members of the SB
- Representative of the stakeholders (COLAF, AFCOS, GdF, …)

What to do

Scientific board (SB):

–Statisticians and data scientists;
–Experts in business management;
–Expert IT
–Expert in EU budget

How to do it


EB
SB
Stakeholders

Data management

- It is not possible to accomplish any of the tasks previously discussed without having access to data;
- At least two difficulties:
  - Privacy and confidentiality issues;
  - Data have high dimensionality and require hardware with adequate computing power

Thank you!