CRISEI
Centro di Ricerca Interdipartimentale in Sviluppo Economico e Istituzioni

Discussion Paper Series

# *Estimation of criminal populations using administrative registers in the presence of linkage errors: enhancing the quality of national accounts*

**Antonella Baldassarini, Valentina Chiariello, Tiziana Tuoto**

**Discussion Paper No. 02**
**February 2018**

**ISSN: 2280-9767**

CRISEI - Università di Napoli - Parthenope

# Università degli Studi di Napoli - Parthenope

# CRISEI

***Estimation of criminal populations using administrative registers in the presence of linkage errors: enhancing the quality of national accounts***

**Antonella Baldassarini*, Valentina Chiariello*, Tiziana Tuoto***

* ISTAT

Comitato Editoriale

Carlo Altavilla,
Maria Rosaria Carillo,
Floro Ernesto Caroleo,
Marco Esposito,
Luigi Moschera,
Oreste Napolitano,
Alessandro Sapio

Via Generale Parisi, 13 - 80132 –
Napoli (Italy) Tel. (+39) 081 547 42 36
Fax (+39) 081 547 42 50
URL: http://www.crisei.uniparthenope.it/DiscussionPapers.asp

# Estimation of criminal populations using administrative registers in the presence of linkage errors: enhancing the quality of national accounts[1]

Antonella Baldassarini Valentina Chiariello Tiziana Tuoto
ISTAT

**Abstract**

The paper describes an approach aimed to improve accuracy and reliability of the labour input engaged in illegal activities using administrative databases. Data refers, in particular, to the alleged crimes for which judicial authority started a criminal proceeding and for which have been enrolled in the registrations of the Public Prosecutor's offices. Administrative databases are used like potential registers of the known criminals. The paper presents an original methodology for measuring how many criminals have missed by the justice system but active in the illegal market based on a capture-recapture approach and the Zeltermann estimator.

**Keywords:** illegal activities, labour, administrative archive, Zeltermann estimator, linkage errors

[1] We thank Maria Giuseppina Muratore and Alessandra Capobianchi for the work provided in identifying and processing criminal administrative databases and for their helpful suggestions as experts of criminal justice statistics.

## 1. Introduction

According to the European legislation, national accounts have to include illegal activities covering exhaustively the economic transactions occur in the economic system[2]. A complete coverage of economic transactions is an important aspect of the quality of national accounts and of the gross domestic product indicator (GDP). Estimates on illegal activities are mainly based on administrative databases or on information that nevertheless present a considerable margin of uncertainty.

The inclusion of specific illegal activities (in particular, production and trafficking of drugs, smuggling of alcohol and tobacco and prostitution services) in the national accounts is a decision that has been taken at European level and makes operational a principle present in the national accounts regulations already since 1995. In Italy, the national accounts estimates, comprehensive of the illegal activities have been disseminated since September 2014 and rebuilt up to the year 1995.

The basic data used for the estimation are of a public nature but do not come from official statistics that, up until now, have not dealt with the direct measurement of these activities. In general, since illegal activities are practiced by individuals with strong incentives to conceal their involvement, both as producers and as consumers, estimates are affected by a margin of error much higher than those characterize other components of GDP.

Estimation of illegal value added is based on methodological approaches shared between statistical offices. The approaches refer to demand indicators when the information sources allow it; in the above case, the estimates are mainly based on information concerning the end users of the illegal good or service and their consumption behaviours. In other cases, supply indicators are used to estimate the value of production starting from information on the goods seized or on the production units involved. The labour input measure is obtained indirectly from estimates of value added, such as in the case of drugs and smuggling of cigarettes, or from expert estimates such as prostitution.

The paper describes a methodology aimed to improve accuracy and reliability of labour input estimates for illegal activities using an administrative database of the Ministry of Justice available for several years. The source refers, in particular, to the alleged crimes for which

[2] Regulation (EU) No 549/2013 of the European Parliament and of the Council of 21 May 2013 on the European system of national and regional accounts.

judicial authority started a criminal proceeding and for which have been enrolled in the registrations of the Public Prosecutor's offices. The methodological approach is able to solve limitations related to the administrative nature of the data source, that is grossing up to unknown population, and the record linkage errors; these last are due to the incomplete personal code available that refers, for privacy motivations, only to soft personal information.

## 2. Data sources

The Ministry of Justice provides to Istat annually data on alleged crimes for which the judicial authority started a criminal proceeding and which have been enrolled in the registers of the Public Prosecutor's offices. Crimes that are registered in the criminal registers of the Public Prosecutor's offices represent the first step of official knowledge about the proceeding. These data are provided to Istat without unique identifier for criminals, only soft identifier as date, place of birth and gender are available. Based on this information, the crime authors are identified and followed in a specific time span. In this way, the administrative source can be considered as a list of criminals with the count (i.e. the number of times) that they appear in the Prosecutor's offices registers. The administrative register also provides some characteristics of the criminals and the crime acts, like age in the time of crime, nationality, the association with other criminals and other crimes done. This information can be exploited to explain heterogeneity in the individual behaviours.

On the other hand, the lack of unique identifiers and the risk of false links due to the use of soft identifiers in linkage procedure have to be solved. The labour force survey (LFS) has been used as additional data source where complete identifiers are available. Information refers to legal workers under the hypothesis that false match rates are similar to those of illegal workers. The false match rate has been estimated considering the occurrence of coincidence on the birth date, gender and place of birth due to chance on distinct individuals in the LFS. The covariate ‘age’ goes to 18 until 80 years old. The covariate ‘gender’ indicates the sex of the criminals. The covariate ‘nationality’ indicates if the criminal is Italian or foreigner. The covariate ‘other crimes’ indicates if the author has also pending denunciation for other crimes. The covariate ‘association’ indicates if the offender has committed the crime in association with other people.

## 3. International experiences and the Zelterman estimator

Hidden criminal population have been estimated since a long time with capture-recapture methods and the Zelterman estimator. International experiences to which the present work refers are the following: Rossmo and Routledge (1990) that estimated the size of criminal populations of migrating fugitives and for street prostitute with capture-recapture analysis; Van Der Heijden et al. (2003) that estimated the size of criminal population of drunk drivers and persons who illegally possess firearms using a truncated Poisson regression model and build the dependent variable with capture-recapture method from Dutch police records; Bouchard and Tremblay (2005) that employed a capture-recapture method to determine the size of hidden population of drug dealer and consumer in Quebec; Rossi (2013) that published an estimation of the hidden population of drug dealers and consumer population employing different methods (including the Zelterman estimator). The results of the above last study are not so far from the results here presented even if different data sources are employed for which it must be considered the risk of linkage error.

According to the approach here presented, the 2006-2012 administrative databases available are considered as a list of individuals from the potential population of criminals; annual databases make possible to count more than one time each individual, even if with some uncertainty due to the risk of false links. Furthermore, some criminals are not observed at all, so the list can be incomplete and show only part of the population. In this framework, several methods have been studied for estimating the population size, where the question is mainly how many individuals are missed by the register.

Shortly, the register's counts are considered to come from a zero-truncated Poisson distribution: the observed counts $f_1 + f_2 + f_j + \cdots = n_{obs}$ give the size of the list but the frequency of not observed units $f_0$ is unknown. The size of the total population is $N = n_{obs} + f_0$.

Given that all units of the population have the same probability $P(Y > 0) = 1 - P(Y = 0)$ of being included in the list, the population size N can be estimated by means of the Horvitz-Thompson estimator:

$$\widehat{N} = \sum_{i=1}^{n_{obs}} \frac{1}{1-P(Y=0)} = \frac{n_{obs}}{1-\exp(\lambda)} \quad (1)$$

Where $i$ are the observed individuals and $\lambda$ is the unknown parameter of the Poisson distribution. Clearly, $\lambda$ can be estimated with maximum likelihood under the assumption of a homogeneous truncated Poisson distribution (Böhning, van der Heijden, 2009, van der Heijden et al. 2003). In alternative, some different estimators have been proposed. For

instance, the Zelterman estimator only uses the first two counts so it is less sensitive to model violations than the estimator that assumes homogeneous Poisson distribution for the entire range of frequencies fj. Indeed, Zelterman (1988) argued the Poisson assumption might not be valid over the entire range of possible values for Y but it might be valid for small ranges of Y such as from j to j +1. The original formulation of the Zelterman estimator is based on a property of the Poisson distributions, which also works for zero-truncated Poisson distributions. Zelterman suggested λ can be estimated by using the frequencies closest to the target prediction $f_0$, that is $f_1$ and $f_2$, as:

$$\hat{\lambda}_Z = \frac{2f_2}{f_1} \tag{2}$$

The above estimator is unaffected by changes in the data for counts larger than 2 and this contributes largely to its robustness; this solution seems particularly proper in this application because of the observed count distribution, with debatable high level frequencies, up to $f_{70}$.

The resulting estimator for the population count is:

$$\hat{N}_Z = \frac{n_{obs}}{1-\exp(\hat{\lambda}_Z)} = \frac{n_{obs}}{1-\exp\left(\frac{2f_2}{f_1}\right)} \tag{3}$$

The Zelterman estimator can be adjusted to take into account covariates to explain the observed heterogeneity (Böhning and van der Heijden, 2009). The covariates can be incorporated into the modelling process to include individual heterogeneity in $\lambda_i$, by

$$\lambda_i = 2\exp(\beta^T x_i), \tag{4}$$

where $x_i$ is the vector with covariate values and $\beta$ is the corresponding parameter vector. In this work, a generalized Zelterman estimator can be derived for the population size N:

$$\hat{N}_{GZ} = \frac{n_{obs}}{1-\exp(2\exp(\beta^T x))}. \tag{5}$$

In this application, the available covariates refer to socio-demographic characteristics of the criminals (that is, gender, age, nationality) and features of the criminal activities, that is the criminal acts in association with other people (As) and the involvement in other kinds of crimes during the reference period (OC). A model selection can be applied in order to select the proper covariates according to the parsimony's principle, identifying, if necessary, different models for each kind of crime: for instance, first results on drug market show that the variable 'the criminal acts in association with other people' is the most relevant one in explaining heterogeneity in capture probabilities.

Finally, the Zelterman estimator, both the simple one, formula (3), and in the presence of covariates, formula (5), can be adjusted to avoid bias related to the potential false linkage errors caused by the lack of strong identifiers. In fact, due to false linkage errors, the observed counts, say $f_j^*$, can be inflated or deflated compared to the true values $f_j$. One can assume that the relationship between the observed counts and the true ones can be explained by the false linkage errors and in this way it is possible to further adjust the Zelterman estimator. Moreover, confidence intervals of the proposed estimates are provided, under the assumption of known linkage errors (Bohning, 2010). Extensions to the case of estimated linkage errors can be studied.

## 4. Results and linkage errors

The known population involved in the crimes under observation is shown in Table 1. Data are referred to the estimated number of crime authors in presence of not univocal identification number. They have been obtained aggregating basic information for personal data, types of crime and year of the procedure, number of known population involved in the crimes under observation, are referred to the estimated number of crime authors in presence of not univocal identification number.

**Table 1. Number of crimes per year of the procedure's beginning**

| Year | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|---|
| Drug trafficking | 35,486 | 38,114 | 40,537 | 41,114 | 37,573 | 37,034 | 34,100 | 36,584 | 34,964 |
| Exploitation of prostitution | 2,784 | 2,929 | 3,193 | 3,030 | 3,109 | 2,955 | 2,831 | 2,717 | 2,740 |
| Smuggling | 1,883 | 2,102 | 2,543 | 3,386 | 2,349 | 2,261 | 2,802 | 2,924 | 3,349 |

The number of proceedings enrolled by the Public Prosecutor's Offices is considered as a list of criminals and, using the personal information available, it is possible to count how many times a potential criminal appears on the list. Results are presented only for the population involved in drug trafficking that records the highest number of crimes data referred to a benchmark year 2012.

The presence of counts of frequencies for drug trafficking is high order (up to 70) and it is confirmed also for the other years. The other crimes present frequencies quite lower than drugs, e.g. the highest frequency for prostitution exploitation is around 10 while the highest frequency for smuggling is around 30. This may be due to the fact that having carried out multiple crimes the judiciary has opened more proceedings for individual crimes or a defect

of the dataset. However, in this case, the use of a robust estimator like the Zelterman seems to be recommendable to reduce the sensitivity of the results with respect to the changes in the data for counts larger than 2.

Administrative records also provide certain characteristics of criminals and criminal acts, such as age at the time of the crime, nationality, association with other criminals and other crimes committed. Information can be exploited explaining heterogeneity in individual behaviours; for the above reason, the information has been used in the model like covariates. Furthermore, in this way it is possible to obtain estimates of the observed and unobserved criminal population distinct from the characteristics considered. Covariates help in better explaining the heterogeneity of the parameters (G=gender, A=age, N=nationality, OC=other crimes, As= association). Covariates for drugs related crimes can be summarized as follows: the covariate 'age' shows that most criminals are concentrated in the age range to 20 until 50. The covariate 'gender' indicates that criminals are predominantly males. The covariate 'nationality' indicates that criminals obviously are more Italian than foreigner but foreigners are a consistent number.

**Table 2: Counts for covariates for drug crimes in 2012**

| Covariate | f1 | f2 | Counts |
|---|---|---|---|
| **Female** | 2,206 | 243 | 2,644 |
| **Male** | 26,053 | 3,296 | 31,456 |
| **_<30 years** | 14,881 | 1,889 | 17,909 |
| **30-50 years** | 11,920 | 1,466 | 14,440 |
| **>50 years** | 1,458 | 184 | 1,751 |
| **Italians** | 18,421 | 2,221 | 22,081 |
| **Foreigners** | 9,838 | 1,318 | 12,019 |
| **Not-involved in other crimes** | 22,627 | 2,563 | 26,813 |
| **Involved in other crimes** | 5,632 | 976 | 7,287 |
| **Act alone** | 16,483 | 1,128 | 17,916 |
| **Act in association** | 11,776 | 2,411 | 16,184 |

Moreover, most of the criminals do not have denunciations about other types of crimes, even if the denunciations for other crimes are not few; the counts are finally almost divided in half

between those who acted alone and those who acted in association with other people. This situation is because many of the subjects denounced for drug related crimes are drug dealers who are predominantly young, many of whom are foreigners, affiliated to national organizations that manage drug trafficking.

Table 3 show the covariates that most affect the dependent variable and the different models to take into account for drug related crimes. The analysis highlights that crimes related to drugs involve mainly men that have been involved in other crimes in association with other people; the above results strength the thesis that the drug related crimes are the typical crimes done within criminal organization. The most relevant covariate on drug crimes is "crime in association with others".

**Table 3: Models for drugs related crimes in 2012**

| Model | Akaike | $G^2$ Test | N | C.I. |
|---|---|---|---|---|
| **G+A+N+OC+As** | 21,254.14 | Accept remove A | 191,149 | 182,575 – 199,723 |
| **G+N+OC+As** | 21,252.15 | Accept remove N | 191,149 | 182,575 – 199,723 |
| **G+OC+As** | 21,258.6 | Reject remove none | 190,706 | 182,178 – 199,234 |
| **G+OC** | 22,107.96 | Reject remove OC | 157,236 | 151,991 – 162,481 |
| **G** | 22,208.78 | Reject remove G | 154,083 | 149,116 – 159,050 |
| **OC** | 22,107.7 | Reject remove OC | 157,148 | 151,913 – 162,383 |
| **Null** | 22,210.81 | | 153,905 | 148,957 – 158,853 |
| **As** | 21,318.63 | Reject remove As | 188,228 | 179,915 – 196,541 |

As previously stated, it is assumed that linkage errors, in particular false linkage, may affect the observed counts, then the relationship between observed counts and true ones via the linkage errors is modelled.

Moreover, it is evaluated the linkage errors on a set of data related to people working on legal activities, i.e. the Labour Force Survey sample. Person identifiers are known for these data, as well as demographic attributes used to recognize the individuals in the potential criminal register. Comparing the results of linkage performed via the personal identifiers with the

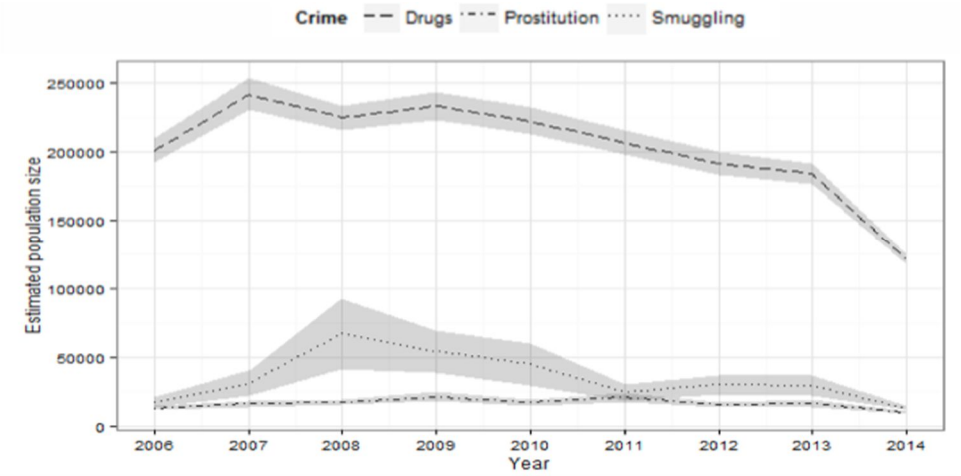
results from the linkage based on soft attributes, the probability of being linked in the criminal register is calculated.

As expected, the frequency of matches purely increases when increasing the size of the considered records. A random sample from the LFS of the same size of criminal population for each class of investigated crimes has been drawn so to measure the frequency of matches of the soft identifiers in similar conditions. The linkage errors appear negligible for population size similar to those involved in prostitution exploitation and smuggling. On the contrary, with numbers like the crimes related to drugs, it results that the frequency of matches by chance is about 1.4%. Moreover, it is almost doubled for foreigners compared to Italians (2.72% and 1.28% respectively). These quantities have been used to adjust the estimates of the criminal population size for drugs crimes, according to the methodology illustrated in section 3. For instance, Table 4 reports the adjusted and naive estimates for crimes related to drugs for some models considered in the previous paragraph.

**Table 4: Comparison of linkage error adjusted estimates for drugs related crimes in 2012**

| Model | Nationality | Ignoring linkage errors | | Adjusting linkage errors | |
|---|---|---|---|---|---|
| | | N | C.I. | N | C.I. |
| As | Total | 188,228 | 179,915 – 196,541 | 191,346 | 182,845 - 199,847 |
| | Italians | | | 127640 | 120,494 – 134,786 |
| | Foreigners | | | 64849 | 60,091 – 69,607 |

**Figure 1. Illegal population by type of crime. Years 2006-2014**



Final results on the illegal population involved in drug trafficking, exploitation of prostitution and smuggling are shown in Figure 1.

## 5. Conclusions

Illegal activities for their nature are difficult to measure as people involved have obvious reasons to hide these activities. As consequence, no official statistics exist while administrative data sources and statistical techniques for recording illegal population and activities are generally not homogenous . The methodology based on the use of the Zelterman estimator aimed to solve grossing up to unknown population and the record linkage errors; linkage errors are due to the incomplete personal code available, referred for privacy motivations only to soft personal information. The extension of the Zelterman estimator to the presence of linkage errors is an innovation applicable to other applications in presence of uncertainty in the unit identification. It provides estimates of the population of illegal authors who perform crimes related to drugs, prostitution exploitation and smuggling.

Considering the difficulties of this kind of estimation and the inaccuracy of the data sources available, the analysis of administrative data here presented seems enough accurate unless errors due to non-statistical nature of the collected data. It is a first step in the application of a statistical approach to provide estimates of the illegal population, in particular for crimes with a high level of recidivism.

**Essential references**

Blumstein, A. (1986), Criminal Careers and "Career Criminals". Vol. 2. National Academies.

Bohning, D. and Van Der Heijden, P. G. M. (2009), A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations, Annals of Applied Statistics, 3, pp. 595-610.

Bouchard M. and Tremblay P. (2005), Risks of arrest across drug markets: A capture-recapture analysis of hidden dealer and user populations, Journal of drug issues, 35(4) pp. 733-754.

Rey G.M., Rossi C, Zuliani A. (2011), Il mercato delle droghe: dimensione, protagonisti e politiche, Marsili editori, Venezia.

Rossi, C. (2013), Monitoring the size and protagonists of the drug market: Combining supply and demand data sources and estimates. Current drug abuse reviews, 6(2), pp. 122-129.

Rossmo D. K., Routledge R. (1990), Estimating the size of criminal populations, Journal of Quantitative Criminology.

Van Der Heijden, P.G.M., Cruyff M., and Van Houwelingen H.C. (2003), Estimating the size of a criminal population from police records using the truncated Poisson regression model, Statistica Neerlandica, 57(3), pp. 289- 304.

Zelterman D. (1988), Robust estimation in truncated discrete distributions with application to capture-recapture experiments, J. Statist. Plann. Inference 18(2), pp. 225-237