Estimating Criminal Populations from Administrative Registers

Antonella Baldassarini[1] | *ISTAT, Rome, Italy*
Valentina Chiariello[2] | *University of Naples, Naples, Italy*
Tiziana Tuoto[3] | *ISTAT, Rome, Italy*

Abstract

This study proposes a methodology for estimating the hidden criminal population working in markets of drug trafficking, prostitution exploitation and smuggling in Italy during the period 2006–2014. These estimates represent the first step of a wide procedure that has the final objective of measuring the economic flows of illegal transactions in national accounts. We exploit administrative registers coming from the Ministry of Justice, and consider these registers as lists of potential criminals. Unique codes for denounced criminals are not available, limiting so far its exploitation at micro level. This drawback has been overcome in this work by proposing an adjustment of the Zelterman estimator that accounts for the potential linkage errors caused by the lack of exact unique identifiers in the dataset. We obtain yearly estimates of the population size of criminals including also the unknown population, for the crimes of drug trafficking, prostitution exploitation and smuggling during the period 2006–2014.[4]

Keywords

Illegal economy, population size estimation, Capture-recapture, Zelterman estimator, linkage errors

JEL code

C13, E26, P41, K42

INTRODUCTION

This study seeks to estimate the hidden criminal population working in the drug markets, prostitution and smuggling of cigarettes in Italy during the period 2006–2014. This research aims to assess the size of illegal markets considered a substantial part of the illegal economy and more widely fits the field of measuring the flow of illegal proceeds to adjust GDP. According to European regulation, national accounts aggregates have to include illegal activities covering exhaustively the economic transactions which occur in the economic system. A complete coverage of economic transactions is an important aspect of the quality of national accounts.

The inclusion of illegal activities (in particular, the production and marketing of drugs, alcohol and tobacco smuggling, and prostitution) in national accounts estimates is a decision that has been taken

[1] ISTAT, Via Depretis 74/b, 00184 Rome, Italy. Corresponding author: email: anbaldass@istat.it, phone (+39)0646733148.

[2] University of Naples "PARTHENOPE", Via Generale Parisi 13, 80132 Naples, Italy.

[3] ISTAT, Via Depretis 74/b, 00184 Rome, Italy.

[4] This paper is based on contribution at the *European Conference on Quality in Official Statistics (Q2018)* in Krakow, June 2018, Poland.

at the European level[5] and implemented by Member States following Eurostat recommendations in terms of methodological approach, quality and reliability of data sources, identification and solution of double counting.

Illegal activities for their nature are difficult to measure as people involved have obvious reasons to hide these activities. As a consequence, data sources and statistical techniques for measuring illegal economic activities are generally not homogenous and standardized.

Italy is considered one of the key countries in Europe for drug trafficking that has long since reached a transnational dimension both because of the geographic position in the Mediterranean Sea and presence of criminal organizations. The big size of this market does not concern production but only import. In Italy, prostitution is legal while it is a criminal offense to organize and exploit prostitutes. Italy adopts an approach towards prostitution of the neo-abolitionist type: outdoor prostitution is neither prohibited nor regulated, while the indoor kind is forbidden only in brothels. In essence, it is not a crime to offer paid sexual services or buy sexual services for a fee. Activities typically associated with prostitution, such as exploitation, recruitment and abetting are instead punished by law. Smuggling activities identify the violation of tax provisions relating to the manufacture, trade and consumption of products subject to the payment of a manufacturing or consumption tax. Smuggling activity goes often along with the counterfeit of cigarettes; the counterfeit of cigarettes is the production of not authentic tobacco products using a trademark without the authorization of the owner. Italy is mainly a territory of final distribution of illegal cigarettes than of transit.

Actually, the Italian National Institute of Statistics (ISTAT) estimate applies both a demand-side and supply-side approaches. Estimating the value added and other aggregates on demand indicators means to use information regarding the final users of the illegal goods or services and their consumption behaviour. In other cases, supply indicators are used, estimating the value of production from information on the production units involved or the goods seized by authorities.

Illegal market of drugs of trafficking is estimates by demand side, prostitution and smuggling of cigarettes by supply side. Trying to provide a complete coverage of the size of the above illegal markets, we calculate the hidden population of illegal authors for each of these crimes. Hidden criminal populations have been estimated since a long time with capture-recapture methods. Rossmo and Routledge (1990) estimate the size of criminal populations of migrating fugitives and street prostitutes with Capture-recapture analysis. Collins and Wilson (1990) estimate the size of the criminal population of automobile thief in Australia applying the Zelterman method. Van Der Heijden et al. (2003a) estimate the size of criminal population of drunk drivers and persons who illegally possess firearms using a truncated Poisson regression model and building the dependent variable with capture-recapture method from the Dutch police records. In order to estimate the risk of being arrested as a drug dealer and a consumer, Bouchard and Tremblay (2005) employ a capture-recapture method to determine the size of these two hidden populations in Quebec (Canada). Mascioli and Rossi (2008) estimate the drug consumers employing a capture-recapture method. Rossi (2013) estimates the hidden population of drug dealers and consumers employing different method in order to measure the market size from demand and supply side.[6] The results of this study are not so far from ours. The paper describes an approach aimed to improve the accuracy and reliability of the labour input estimates for illegal activities using an administrative database of the Ministry of Justice, available for the period 2006–2014. The source refers, in particular, to the alleged crimes for which judicial authority started a criminal proceeding and which have been enrolled in the registrations of the Public Prosecutor's

[5] The new ESA 2010 regulation states that national accounts data should be subject to assessment according to the quality criteria set out in Article 12(1) of Regulation (EC) No. 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics. One of the criteria established by Eurostat for national accounts estimates is the accuracy and reliability of the estimates annually provided according the ESA 2010 transmission programme.

[6] See also Rey, G. M. et al. (2011).

offices. It is possible to consider the above source as a potential register of the known criminals: in this source it is possible to observe an individual more than once, if he/she is accused of committing more than one offence. However, it is unknown the number of individuals not observed by the justice system even if active in the illegal markets, so units that actually belong to the target population. The list from the Public Prosecutor's offices is therefore incomplete, as it allows us to count only a part of the interest population. This is the key question of this study. To solve this point, we assume that the counts from the Public Prosecutor's offices come from a zero-truncated Poisson distribution and a capture-recapture method with the Zelterman estimator is applied to estimate the number of criminals not observed by the judicial system. In addition, in our context, a complication occurs due to the lack of personal identifiers in the registers, caused by privacy motivations: indeed, only soft identifiers as gender, date and place of birth of the suspected criminals are available to us. In this case, it is possible that the linkage, carried out to retrace and count how many times the same individual appears in the Public Prosecutor's office lists, can be compromised. Intuitively, one can expect that some false matches (that is, false positive) may occur just because some people happen to have the same birth date, gender and place of birth. The reliability of matches is then examined by considering the occurrence of matches purely by chance. This work proposes indeed an adjustment to the Zelterman estimator in order to take into account the linkage errors. Moreover, the generalized estimator allows including covariates in the estimation of the hidden population size. The proposed estimator can be applied in different situations in which the linkage results are subject to uncertainty. For instance, in this case we only assume false links due to the fortuity of sharing the same soft identifiers; however, the same methodology can be applied in the presence of linkage errors due to inaccuracies in the matching variables.

The methodology resolves two limitations related to the nature of the source: firstly the identification of the number of the known persons involved in crimes through the correction of the criminals overlapping; secondly the identification of the proper estimator for grossing up the correct number of criminals to calculate the hidden population of illegal workers in illegal markets of drug, prostitution and smuggling to give a whole dimension to the phenomenon, identifying both the potential known and unknown criminals.

It is worth noting that our analysis deals with an administrative list of denounced crimes, referring suspected subjects in drug trafficking, prostitution exploitation and smuggling, with the aim of estimating the unknown size of people involved in the abovementioned crimes. Our estimates of the hidden criminal population working in markets of drug, prostitution and smuggling in Italy during the period 2006–2014 are coherent with the information coming from other sources.

The paper is structured as follows: in section two we describe the data, in section three we propose the methodology employed in the estimation; in section four we show the results and finally in section five we provide some concluding remarks.

1 THE DATA

This study uses annually official data provided to Istat by the Ministry of Justice regarding the 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 suspected criminals, only soft identifier as date, place of birth and gender are available. Based on this information, the crime perpetrators are identified and followed in a specific time span. In this way, the administrative source can be considered as a list of potential criminals with the count (i.e. the frequency) that they appear in the Prosecutor's offices registers. In the list we can observe individuals who are charged 1, 2, 3, ... k times, however, we cannot observe units not recorded in the Justice system. Statistically, these data can be considered as coming from a count distribution truncated at zero. The administrative register also provides some characteristics of the denounced subjects and the crime acts, like age at the moment of the crime, nationality, the association with other subjects 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. The 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.

Aggregating for personal data, kinds of crime and the year of the proceedings, we observe counts for the three considered crimes as in Table 1.

Table 1 Observed counts for the three crimes of interest by the year of the proceedings

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014
Drugs	35 486	38 114	40 537	41 114	37 573	37 034	34 100	36 584	34 964
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

Source: Our elaboration on the registers of the Public Prosecutor's offices data

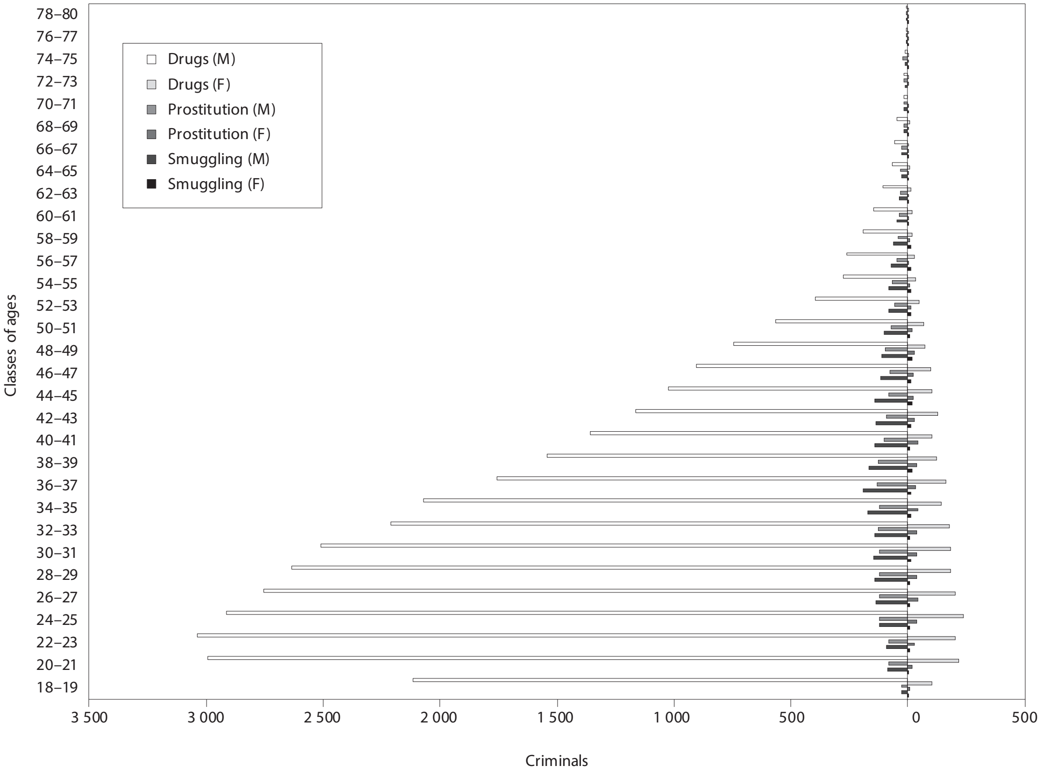
Table 1 shows the yearly number of denunciations for each type of crime and how the number of denunciations changes in time and among types of crime. It is clear that the yearly counts are quite similar for the considered crimes. For this reason, in the rest of the paper, sometimes we just show the results of a single reference year, the 2013. In addition, we decided to show the results of the crime of drug trafficking because it records the highest number of crimes and the methodology for adjusting the linkage error is more effective.

Figure 1 describes the three pyramids of age and gender of the observed populations in 2013, collapsed in just one figure in order to make evident the comparison between the different crimes. The figure shows primarily that the denunciations of drug related crimes committed by men are far greater than the denunciations made for other crimes. Prostitution and smuggling related crimes data also show that almost all the crimes are attributable to men. As we expected, the 20–40 age group is the one with the highest number of denunciations for all the three type of crimes, but in drug trafficking it is more evident than in others.

In this study the considered covariates are gender (G), age (A), nationality (N), other crimes (OC), association (As). We considered these covariates because they can help in better explaining the heterogeneity of the parameter. The covariate "age" goes to 18, the year of the majority age in Italy (that is also the age from which one can go to prison), until 80, the year that we considered possible to carry out a job. The covariate "gender" indicates the sex of the suspected criminals. The covariate "nationality" indicates if the subject is Italian or foreigner. The covariate "other crimes" indicates if the author has also one or more pending denunciation for other crimes. The covariate "association" indicates if the offender has committed the crime in association with other people.

Table 2 reports the covariates for drug crimes in 2013: as already shown by the pyramids in Figure 1, males are far more than females, most subjects are concentrated in the age group under the age of 50. Italians are only twice than foreigners. Moreover, most of the units do not have denunciations about

Figure 1 Pyramids of age and gender of illegal populations in 2013



Source: Our elaboration on the registers of the Public Prosecutor's offices data

Table 2 Counts for covariates for drug crimes in 2013

Covariate	Counts
Female	2 732
Male	33 852
≤ 30 years	18 952
30–50 years	15 555
> 50 years	2 077
Italians	23 917
Foreigners	12 667
Not-involved in other crimes	28 397
Involved in other crimes	8 187
Act alone	19 525
Act in association with other people	17 059

Source: Our elaboration on the registers of the Public Prosecutor's offices data 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. We have this situation because many of the subjects denounced for drug related crimes are drug dealers who are predominantly young, many of whom are foreigners but also Italians, affiliated to national organizations that manage drug trafficking.

2 METHODOLOGY

2.1 The Zelterman estimator

As shown in the previous section, the administrative register from the Justice Ministry can be viewed as a list of individuals from the population of criminals, where we are able to count how many times each individual is registered, even if with some uncertainty due to the risk of false links. However, some population members 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 counts are considered to come from a zero-truncated Poisson distribution: according to a standard formulation, consider a population of size N and a count variable Y taking values in the set of integers {0, 1, 2, 3, ...}. In this study Y represents the number of criminal proceedings a person has been enrolled in the registration on the Public Prosecutor's offices in the reference time. Denote with $\{f_0, f_1, f_2, ...\}$ the frequency with which a 0, 1, 2, 3, ... occurs in this population.

Since a unit is observed only if Public Prosecutor's offices start a criminal proceedings against him/her, the subject will only be observed if there has been a positive number of proceedings with the justice institution, whilst y = 0 will not be observed in the list. Hence the list reflects a count variable truncated at zero that we denote by Y_+. Accordingly, the list has observed frequencies $\{f_1, f_2, ...\}$ but the frequency f0 of zeros in the population is unknown. The size of the list is not N but nobs, where $N = n_{obs} + f_0$ is the unknown size of the population.

The distributions of the untruncated and truncated counts are connected via:

$$P\,(Y_+ = j) = P\,(Y = j)\,/\,(1 - P\,(Y = 0)),$$

for $j = 1, 2, 3, ...$ For example, if Y follows a Poisson distribution with parameter λ so that:

$$P\,(Y = j) = Po\,(j \setminus \lambda) = exp\,(-\lambda)\,\lambda^j\,/\,j!, \tag{1}$$

for $j = 0, 1, 2, 3, ...$ then the associated distribution of Y_+ is given by:

$$P\,(Y_+ = j) = Po_+(j|\lambda) = \frac{\exp\,(-\lambda)}{1 - \exp\,(-\lambda)}\,{}^{\lambda^j}\!\!/_{j!}\,, \tag{2}$$

with $j = 1, 2, 3, ...$

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

$$\hat{N} = \sum_{i=1}^{n_{obs}} \frac{1}{1 - P(Y = 0)} = \frac{n_{obs}}{1 - \exp\,(-\lambda)}\,. \tag{3}$$

This approach requires that λ is known and if it is not, it needs to be estimated. Clearly, λ can be estimated with maximum likelihood under the assumption of a homogeneous truncated Poisson distribution. In alternative, some different estimators have been proposed in Van Der Heijden et al (2003a, b) and Bohning et al. (2009). For instance, the Zelterman estimator proposed in Zelterman (1988) 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 f_j. 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:

$$Po\,(j + 1|\lambda) = Po\,(j|\lambda) = \lambda \,/\, (j + 1).$$

So, Zelterman (1988) suggested λ can be estimated as:

$$\hat{\lambda}_j = \frac{(j + 1)f_j + 1}{f_j}\ . \tag{4}$$

Zelterman (1988) also argued to use the frequencies f_j closest to the target prediction f_0, that is f_1 and f_2, this leads to the estimator $\lambda_Z = (2f_2 / f_1$, obtained by (4) for $j = 1$.

This estimator is unaffected by changes in the data for counts larger than 2, 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}, as shown for instance in Table 3.

Table 3 Frequencies of captures for drug crimes in 2013

	Counts		**Counts**
f_1	29 755	...	...
f_2	4 108	f_{41}	1
f_3	1 070	f_{42}	1
f_4	542	f_{43}	1
f_5	275	f_{44}	1
f_6	209	f_{45}	2
f_7	115	f_{51}	1
f_8	107	f_{59}	1
f_9	76	f_{65}	1

Source: Our elaboration on the registers of the Public Prosecutor's offices data

The presence of counts of high order, up to 70, is confirmed also for the other years. The other crimes present frequencies quite lower than drugs, e. g. the highest frequency for prostitution is around 10 and the highest frequency for smuggling is around 30. This may be due to either 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.

The resulting estimator for the population size is $\hat{N}_Z$:

$$\hat{N}_Z = \frac{n_{obs}}{1 - \exp(-\lambda_Z)} = \frac{n_{obs}}{1 - \exp(-2f_2/f_1)} \quad . \tag{5}$$

2.2 The Zelterman estimator with covariates

The Zelterman estimator can be extended so to take into account covariates to explain the observed heterogeneity as in Bohning et al. (2009). Indeed, in most applications, the assumption of homogeneous λ is not realistic while the register contains, together with the counts Y, also some information about the individual characteristics.

The covariates can be incorporated into the modeling process, by:

$$\lambda_i = 2\, exp(\boldsymbol{\beta}^T \boldsymbol{x}_i),$$

where $\boldsymbol{x}_i$ is the vector with covariate values including a constant, and $\boldsymbol{\beta}$ is the corresponding parameter vector.

Accordingly, a generalized Zelterman estimator can be derived for the population size N:

$$\hat{N}_{Z_G} = \sum_{i=1}^{n_{obs}} \frac{1}{1 - \exp(-\hat{\lambda}_i)} = \sum_{i=1}^{n_{obs}} \frac{1}{1 - \exp(-2\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)} \, . \tag{6}$$

In this application, the available covariates refer to both socio-demographic characteristics of the potential criminals (that is, gender, age, nationality) and features of the criminal activities (that is, the subject acts in association with other people, the subject is involved in other kinds of crimes during the reference period). A model selection can be applied so to select the proper covariates according to the principle of parsimony, identifying, if necessary, different models for each kind of crime.

This generalized formulation of the Zelterman estimator can be seen as a maximum likelihood estimator (MLE): indeed, as demonstrated in Bohning et al. (2009), a Poisson distribution with parameter λ constrained to values $Y = 1$ and $Y = 2$ yields a binomial distribution with parameter:

$$p = (\lambda / 2) / (1 + \lambda / 2) = \lambda / (2 + \lambda). \tag{7}$$

So the associated likelihood L for the event $Y = 2$ with parameter $p = \lambda / (2 + \lambda)$ is:

$$L = \prod_{i=1}^{f_1+f_2} (1 - p)^{y_i - 1}\, p^{y_i} = (1 - p)^{f_1} p^{f_2} \, . \tag{8}$$

The binomial likelihood is maximized for $\hat{p} = f_2 / (f_1 + f_2)$, that is $\hat{\lambda} = \frac{\widehat{2\hat{p}}}{1 - \hat{p}} = 2f_2 / f_1$. A great advantage of considering the Zelterman estimator as a MLE is related to the availability of its variance in a closed form.

2.3 Zelterman estimators in the presence of linkage errors

Sometimes, the identification of the units in the register/list can be affected by errors, for several reasons: typos or missing values in the identifiers, lack of complete information for privacy preserving. The errors in the unit identification can be of two types: false negative, i.e. missing link of records which actually belong to the same unit, and false positive, i.e. false link of records which actually belong to different unit. In many applications of record linkage, that is the set of methods and techniques aiming at identifying the same unit even if differently represented in data sources, it is often easy to reduce the false links, e.g. by using restrictive acceptance criteria. However, this often increases the number of missing links. In many studies of animal populations, based on the recognition of individual animals from natural markings (e.g. natural tags, photographs, DNA fingerprints), as well as in epidemiology studies, the probability of false links is often negligible, due to the caution in linkage procedures and one should only consider the risk of missing true links. In this study, we assume that linkage errors are not generated by inaccuracy in the matching variables, as usual in the literature on record linkage, but they are the results of the unavailability of complete strong identifiers that generates random matches of partial soft identifiers. So, in this application, the risk of missing true links can be assumed as negligible, while we have to take into account the false positives. In fact, data on proceedings from the Public Prosecutor's Offices are available at Istat without the codes for personal identification, i.e. without names and surnames, but only with soft identifiers like the date and place of birth and gender of person involved in the proceedings. In this case, one can suspect that the efficacy of retracing the counts for each individual can be compromised by the lack of either name or a common person identifier. Intuitively, we can expect that some false matches (that is, false positive) may occur just because some people happen to have the same birth date, gender and place of birth. The reliability of matches can be examined by considering the occurrence of match purely by chance due to the occurrences of birth dates, places and gender.

The Zelterman estimator, both the simple one and in the presence of covariates, 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 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 one can be explained by the false linkage errors and in this way it is possible to further adjust the Zelterman estimator.

Assuming false match rate α affects count f_2 in this way:

$$f_2 = (1 - \alpha)\, f_2^*,$$

i.e., only part of the observed f_2^* are true f_2, namely the part for which we do not expect linkage errors.

Consequently, we have:

$$f_1 = f_1^* + 2\alpha f_2^*,$$

$$n_{obs} = n_{obs}^* + \alpha f_2^*,$$

where f_j represents the true value and f_j^* is the observed one, for $j = 1, 2$.

The linkage errors can be considered in the likelihood:

$$log\ L = \sum_{i=1}^{f_1^*+f_2^*} y_i\,(1 - \alpha)\log(p) + (1 - y_i + 2\alpha y_i)\log(1 - p)$$

and the Zelterman estimator adjusted for linkage errors becomes:

$$\hat{N}^{L}_{Z_G} = \frac{n^{L}_{obs}}{1 - \exp\left(-\lambda^{L}\right)} = \frac{n^{L}_{obs}}{1 - \exp\left(-2\exp\left(\boldsymbol{\beta}^{T}\boldsymbol{x}\right)\right)},$$

where n_{obs}^{L} and λ^{L} are, respectively, the linkage-adjusted observed counts and the linkage-adjusted Poisson parameter.

3 RESULTS

In order to estimate the population size of criminals including also the unknown population, for each year and crime, we started from aggregated data in Table 1 seen in the descriptive analysis of the data. Table 4 shows the different models for drug related crimes.

The covariates that most affect the dependent variable is ”Association” but also ”Gender” and ”Other crimes”. Because as evidenced in the Table 4 the Akaike test (AIC) in the model with solely the covariate “Association” is the lowest and very low is also the AIC for the model with the three covariate “Association”, ”Gender” and “Other crimes”. This is an expected result because, as we have seen before, the crimes related to drugs are committed more by men. The model results that they have done other types of crime and that they have made them in association to other people: this strengthens the thesis that the drug related crimes are the typical crimes committed within national organizations that deal with other crimes also. Our results suggest that the estimated criminal population involved in the market of drug for 2013 is around 181 460.

Table 4 Models for drugs related crimes in 2013

Model	AIC	tt^2 Test	N	C.I.
G + A + N + OC + As	24 017	Accept remove A	183 064	175 734–190 394
G + N + OC + As	24 015	Accept remove N	183 061	175 732–190 390
G + OC + As	24 013	Reject remove none	183 057	175 728–190 386
G + OC	24 941	Reject remove OC	154 025	149 318–158 732
G	25 026	Reject remove G	151 782	147 263–156 301
OC	24 941	Reject remove OC	153 937	149 239–158 635
As	24 009	Reject remove As	181 460	174 264–188 656
Null	25 029		151 625	147 122–156 128

Note: G is gender, A is age, N is nationality, OC is organized crime, As is association with other criminals.
Source: Our elaboration on the registers of the Public Prosecutor’s offices data

As stated in section 2.3 we assume that linkage errors, in particular false linkage, may affect the observed counts and we model the relationship between observed counts and true ones via the linkage errors. Moreover, as introduced in section 2, we evaluate the linkage errors on a set of data related to people involved in legal activities, i.e. the labour force sample survey (LFS) carried out by ISTAT in 2014. Personal identifiers are known for these data, as well as demographic attributes used to recognize the individuals in the administrative register. Comparing the results of linkage performed via the person identifiers with the results from the linkage based on soft attributes we assess the probability of being linked by chance in the register. As expected, the frequency of matches purely by chance increases when increasing the size of the considered records. A random sample from the LFS of the same size of suspected criminal population for each class of investigated crimes has been drawn so to measure the frequency of matches by chance of the soft identifiers in similar conditions. The linkage errors appear negligible for population size similar to those involved in prostitutions 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 related crimes, according to the methodology illustrated in section 3.3. For instance, Table 5 reports the adjusted and naive estimates for crimes related to drugs in 2013 for some models considered in the previous paragraph. As expected, the adjusted estimates are only slightly higher than the naive estimates, as the linkage error is still small. It can be observed that the adjusted estimates are in the confidence intervals of the naive estimates and vice-versa. For example comparing the population before and after adjusting for linkage errors of the model with solely the covariate "Association" we can see that the difference is only of 3 049 criminals.

Table 5 Comparison of linkage error adjusted estimates for drugs related crimes in 2013

Model	Ignoring linkage errors		Adjusting linkage errors	
	N	C.I.	N	C.I.
G + N + OC + As	183 061	175 732–190 390	187 518	176 992–198 044
G + OC + As	183 057	175 728–190 386	186 124	178 628–193 620
G + OC	154 025	149 318–158 732	156 687	151 868–161 506
As	181 460	174 264–188 656	184 509	177 148–191 870
Null	151 625	147 122–156 128	154 253	149 642–158 864

Note: G is gender, A is age, N is nationality, OC is organized crime, As is association with other criminals and Null is the null hypothesis.
Source: Our elaboration on the registers of the Public Prosecutor's offices data

The availability of covariates is exploited in order to obtain estimates for subpopulation of interest. For instance, knowing the different linkage errors affecting subpopulation of Italians and Foreigners, we use this covariate to properly adjust the estimates, as shown in Table 6.

Table 6 Comparison of linkage error adjusted estimates for drugs related crimes for sub-populations

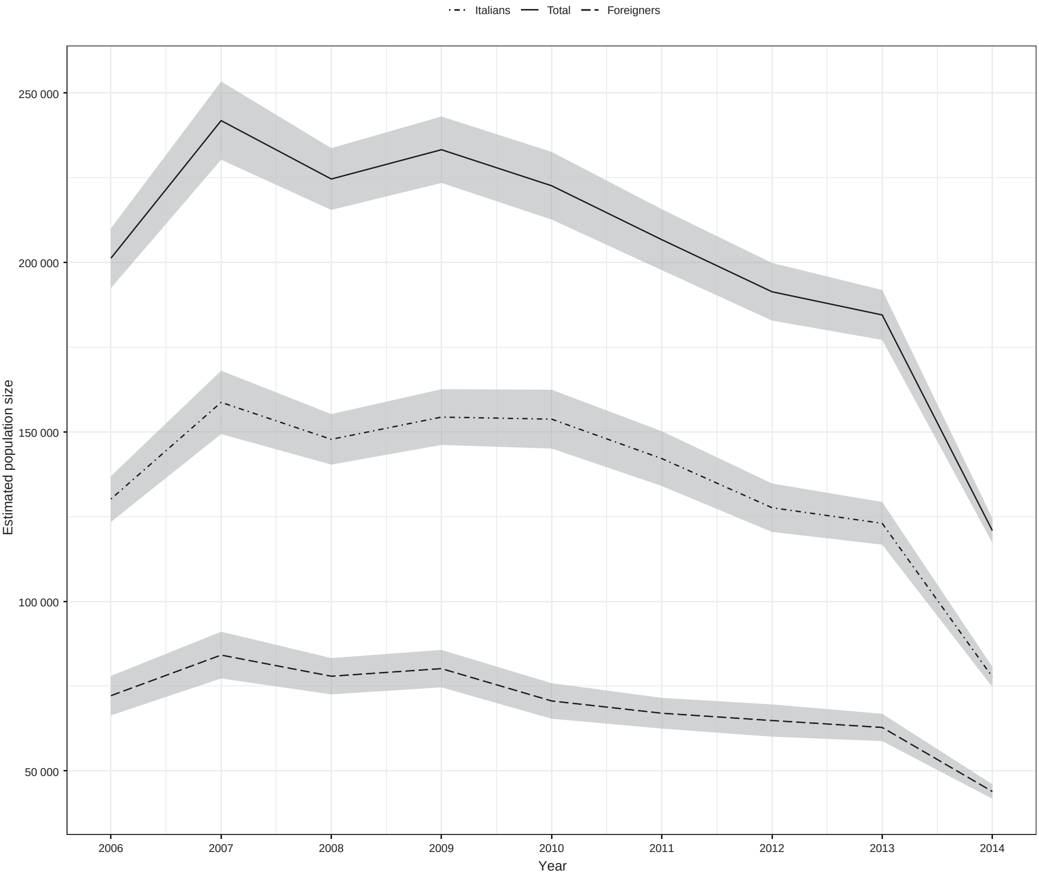
Model	Nationality	Adjusting linkage errors	
		N	C.I.
As	Italians	123 032	116 740–129 324
	Foreigners	62 813	58 779–66 847

Source: Our elaboration on the registers of the Public Prosecutor's offices data

In Figure 2, the population size for drugs is estimated for all the considered period. The confidence intervals of the estimates are also represented. Moreover, since different linkage errors affect Foreigners and Italians, the two subpopulations are adjusted separately, thanks to the available covariance on Nationality.

Finally, Figure 3 shows the population size estimates for all the considered crimes, i.e. drug, prostitution and smuggling.

Figure 2 Estimates for drugs related crimes between 2006 and 2014, for subpopulations



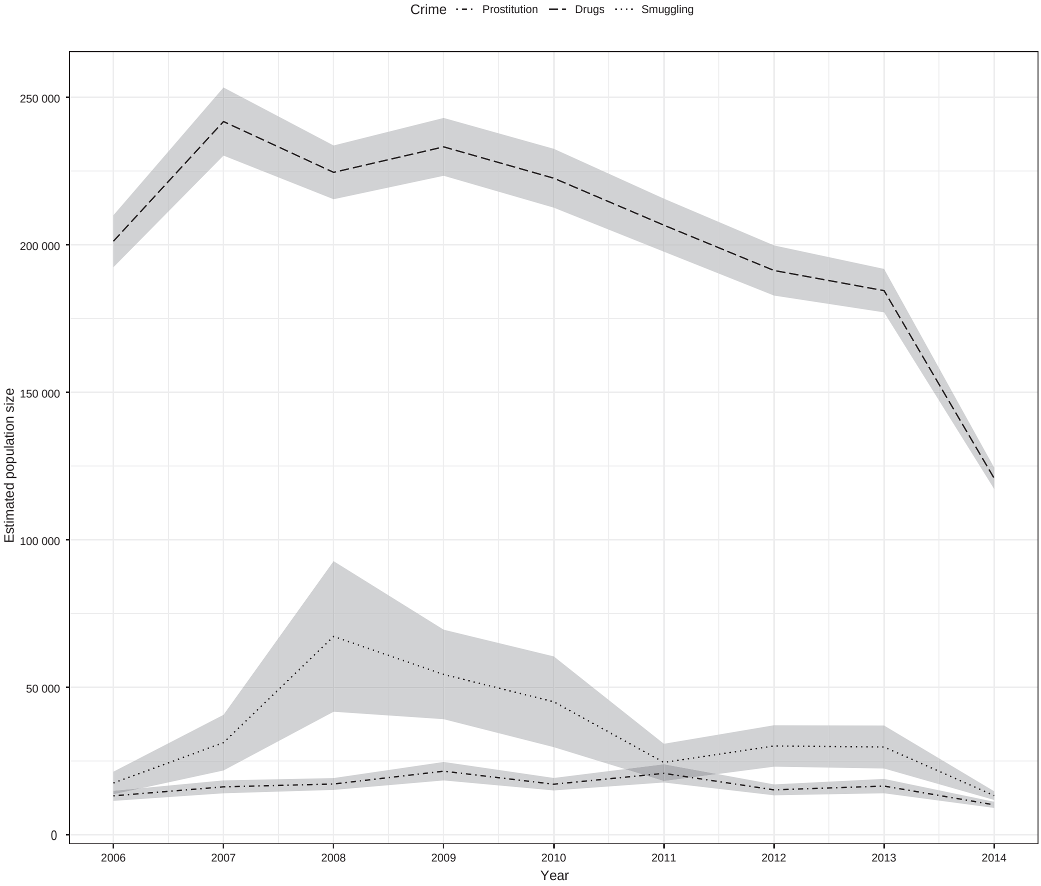
Source: Our elaboration on the registers of the Public Prosecutor's offices data

CONCLUDING REMARKS

The literature on estimating illegal populations with the Zelterman estimator already exists and can be enriched. The innovation added by this work to the literature is the calculation of the illegal population with the Zelterman estimator adjusted for the linkage error, due to the lack of exact identifiers in the dataset.

This study provides an estimate of the population of illegal persons who perform crimes related to drugs, prostitution and smuggling. The analysis observes a set of alleged crimes for which judicial authority started a criminal proceeding and which have been enrolled in the registrations of the Public Prosecutor's offices from 2006–2014. This set is intended as a potential register of the known criminals. To calculate the unknown part of criminals we use the Zelterman estimator. Moreover, due to the absence of an exact identifier for the subjects

Figure 3 Estimates for drug, prostitution and smuggling between 2006 and 2014



Source: Our elaboration on the registers of the Public Prosecutor's offices data

in the data, the Zelterman estimator needs an adjustment for the linkage error. The comparison with other data related to regular workers suggests considering the risk of linkage errors only with the numbers for the drug related crimes. The extension of the Zelterman estimator to the presence of linkage errors can be considered an innovation useful even in other applications subject to the uncertainty in the unit identification. The results of our analysis, as well as providing a number for the illegal population, show that what most affects the increase in the illegal population is the work within a criminal association, having denunciations about other crimes and gender.

Considering the difficulties of this kind of analysis due to the particular field of estimation, the impact that the theme can have on the society, and the inaccuracy of sources in general, our analysis on administrative data seems enough accurate, managing potential errors due to lack of strong identifiers. It could be the first step aiming at providing accurate estimates of the illegal market population in Italy. Further analyses will be dedicated to calculate other populations of illegal actors such as those who commit corruption or who perform other illegal activities considering characteristics specific of the crimes.

References

BOHNING, D AND VAN DER HEIJDEN, PG. M. A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *The Annals of Applied Statistics*, 2009, 3(2), pp. 595–610.

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

COLLINS, M. F. AND WILSON, R. M. Automobile theft: Estimating the size of the criminal population. *Journal of Quantitative Criminology*, 1990, 6(4).

MASCIOLI, F. AND ROSSI, C. Capture-recapture methods to estimate prevalence indicators for the evaluation of drug policies. *Bulletin on Narcotics*, 2010, 60, pp. 5–25.

EUROSTAT. *Regulation (UE) n. 549/2013 of European Parliament and of the Council of 21 May 2013 on the European system of national and regional accounts in the European Union.* 2013.

REY, G. M., ROSSI, C., ZULIANI, A. *Il mercato delle droghe: dimensione, protagonisti e politiche.* Venezia: Marsili editori, 2011.

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

ROSSMO, D. K. AND ROUTLEDGE, R. Estimating the size of criminal populations. *Journal of Quantitative Criminology*, 1990.

VAN DER HEIJDEN, PG. M., CRUYFF, M., VAN HOUWELINGEN, HC. Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica*, 2003a, 57(3), pp. 289–304.

VAN DER HEIJDEN, PG. M., CRUYFF, M., VAN HOUWELINGEN, HC. Point and interval estimation of the truncated Poisson regression model. *Statistical Modelling*, 2003b, 3, pp. 305–322.

ZELTERMAN, D. Robust estimation in truncated discrete distributions with application to capture- recapture experiments. *J. Statist. Plann. Inference*, 1988, 18, pp. 225–237.