Estimation of criminal populations using administrative registers in the presence of linkage errors

Valentina Chiariello Tiziana Tuoto

March 20, 2018

1 Introduction

This study seeks to estimate the hidden criminal population working in markets of drug, prostitution and smuggling in Italy during the period 2006-2014. This research assesses the size of these illegal markets considered a substantial part of the illegal economy and more widely fits into the field of measuring the flow of illegal proceeds in the Italian GDP. According to European regulations, 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 at the European level [1] 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 have long since reached a transnational dimension both because of its geographic position in the Mediterranean Sea and for the presence of criminal organizations. The big size of this market does not concern production but only import. The prostitution is not persecuted neither regulamented but obviously the crime of favouring and exploitation of prostitution is disciplined by the law and persecuted. In Italy this is a crime mostly managed from foreign organizations (e.g. chinese, african and

[1] 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.

east-european). The smuggling of cigarettes is intended as importation and exportation of products that are legal in some other countries. Illegal cigarettes arrive in Italy (especially from Eastern European countries, China and the United Arab Emirates). Indeed, related to these markets there are other economic aspect not to underestimate such as organized crime and corruption of the legal economy by money laundering that came from these activities to facilitate them. Actually, the Italian National Institute of Statistics (ISTAT) estimate proceeds in the Italian GDP of illegal markets of drugs by demand side (data on the number of consumers and quantity consumed), prostitution by supply side (data on the number of prostitutes on the market) and smuggling of cigarettes by supply side too (data on seizures). Trying to provide a more accurate estimate of the size of these illegal markets we calculate the hidden population of illegal authors for each of these crimes. 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 for which have been enrolled in the registrations of the Public Prosecutor's offices. It is possible to consider the above source as a potential register of the known criminals. The main question of this study is how many criminals are missed by the justice system but active in the illegal markets. To solve this point, the counts from the Public Prosecutor's offices can be considered as to come from a zero-truncated Poisson distribution and the Zelterman estimator is applied to estimate the number of criminals not observed by the justice system. In addition, a complication occurs due to the lack of personal identifiers in the registers, due to privacy motivations: indeed, only soft identifiers as gender, date and place of birth of the criminals are available to us. In this case, one may suspect that the linkage, carried out to retrace and count how many times the same individual appears within 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 also an adjustment to the Zelterman estimator in order to take into account the linkage errors. Moreover, the generalised estimator allows including covariates in the estimation of the hidden population size. The proposed estimator is quite clear to understand and easy to implement, as the original one, and it can be applied in different situations in which the linkage results are subject to uncertainty. 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 (due to the incomplete personal code available referred, for privacy motivations, only to soft personal information); 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 the potential known and the unknown criminals. 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 information coming from other sources.

The paper is structured as follows: in the section two we describe the data, in the section three we describe the methodology employed to the estimation in section four we describe our results and in section five we conclude.

2 International experiences and available data

Capture-recapture methods have been used since a long time to estimate various hidden criminal or illegal populations. [8] estimate the size of criminal populations of migrating fugitives and for street prostitute with Capture-recapture analysis. [3] estimate the size of the criminal population of automobile thief in Australia applying the Zelterman method for calculations on the hidden population. [9] 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 Dutch police records. [2] in order to estimate the risk of being arrested for a dealer and a consumer employ a capture-recapture method to determine the size of this two hidden population in Quebec (Canada). [**?**] Mascioli and Rossi (2008) estimate the consumer population of drug employing a capture-recapture method. [7] estimated the hidden population of drug dealers and consumer population employing different method in order to estimate the market size from demand and supply side. The results of this study are not so far from ours.

In this work, we exploit administrative data coming from the Ministry of Justice, recording criminal charges of Public Prosecutor's offices. Due to privacy motivations, we have only soft personal information about the criminals. However, we can consider the administrative source as a list of criminals and by this personal information we are able to count how many times a criminal appears in the list.

The Ministry of Justice provides 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, represents 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 at the moment of the crime, nationality, the association with other criminals and other crimes done. This information can be exploited to explain heterogeneity in the individual behaviours.

From the other hand, the lack of unique identifiers and the risk of false links due to the

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	35486	38114	40537	34964	37573	37034	34100	36584	34964
Prostitution	2784	2929	3193	3030	3109	2955	2831	2717	2740
Smuggling	1883	2102	2543	3386	2349	2261	2802	2924	3349

use of soft identifiers in linkage procedure have been solved by considering an additional data source, which provides information on legal workers, i.e. the labour force survey (LFS - ISTAT 2014) under the assumption that it could provide a population similar to that of illegal workforce. Therefore, the size of the false match error made just because some people happen to have the same birth date, gender and place of birth has been estimated by the occurrence of coincidence on these variables for distinct individuals in the LFS, where complete identifiers are also available.

3 The Zelterman estimator

As shown in the previous section, the administrative registers 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 registers 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 $\{0, 1, 2, 3, \dots\}$. 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 refence time. Denote with $\{f_0, f_1, f_2, \dots\}$ the frequency with which a $0, 1, 2, 3, \dots$ occurs in this population.

Since a criminal is observed only if Public Prosecutor's offices start a criminal proceedings against him/her, the criminal will only be observed if there has been a positive number of proceedings with the justice institution and $y = 0$ 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 $\{f_1, f_2, \dots\}$ but the frequency f_0 of zeros in the population is unknown. The size of the list is not N but n_{obs}, where $N = n_{obs} + f_0$.

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

The distribution of the untruncated and truncated counts are connected via

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

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

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

for $j = 0, 1, 2, 3, \ldots$ then the associated distribution of Y_+ is given as

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

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

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 HorvitzThompson estimator

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

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 [11], [9],[10]. For instance, the Zelterman estimator [12] only uses the rst 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, [12] 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 [12] suggested λ can be estimated as

$$\widehat{\lambda}_j = \frac{(j - 1)f_{j-1}}{f_j}. \tag{4}$$

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 2.

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

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

Table 2: Frequencies of captures for drugs related crimes in 2013

	Counts		Counts
f_1	29755	...	...
f_2	4108	f_{41}	1
f_3	1070	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
...	...	f_{71}	1

3.1 The Zelterman estimator with covariates

The Zelterman estimator can be extended so to take into account covariates to explain the observed heterogeneity [11].Indeed, in most applications, the assumption of homogeneous λ is not realistic while the register contains, togheter with the counts Y, also some information about the individuals characteristics.

The covariates can be incorporated into the modeling process, by

$$\lambda_i = 2\exp(\boldsymbol{\beta}^T \mathbf{x_i}),$$

where $\mathbf{x_i}$ 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

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

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, the criminal is involved in other kinds of crimes during the reference period). A model selection can be applied in order to select the proper covariates according to the parsimonys 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.

This generalized formulation of the Zelterman estimator can be seen as a maximum likelihood estimator (MLE): indeed, as stated in [11], 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

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

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

3.2 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, the set of methods and techniques aiming at identify the same unit even if differently represented in data sources, it is often easier to reduce the false links, e.g. by using more 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, on the contrary, 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 IDs for personal identification, without names and surnames, but only with personal information on date, 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 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

$$f_2 = (1-\alpha){f_2}^*$$

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.
The linkage errors can be considered in the likelihood

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

and the Zelterman estimator adjusted for linkage errors becomes

$$\widehat{N}^L_{Z_G} = \frac{n^L_{obs}}{1-\exp(-\lambda^L)} = \frac{n^L_{obs}}{1-\exp(-2\exp(\boldsymbol{\beta}^T\mathbf{x_i}))} \tag{8}$$

4 Results

Table 1 shows the yearly number of denunciations for each type of crime and how the number of denunciations changes in times and among types of crime. It is clear that the yearly counts are quite similar for the considered crimes, for this reason we decided to 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 it is also possible to apply the methodology for adjusting the linkage error. The tables for the other crimes can be shown in an Appendix.

In order to estimate the population size of criminals including also the unknow population, for each year and crime, the number of proceedings enrolled by the Public Prosecutor's Offices are counted, as shown for example for drug in 2013 in Table 2. 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 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.

4.1 The covariates

The considered covariates are G=gender, A=age, N=nationality, OC=other crimes, As= association. We considered this covariates because they can help in better explaining the

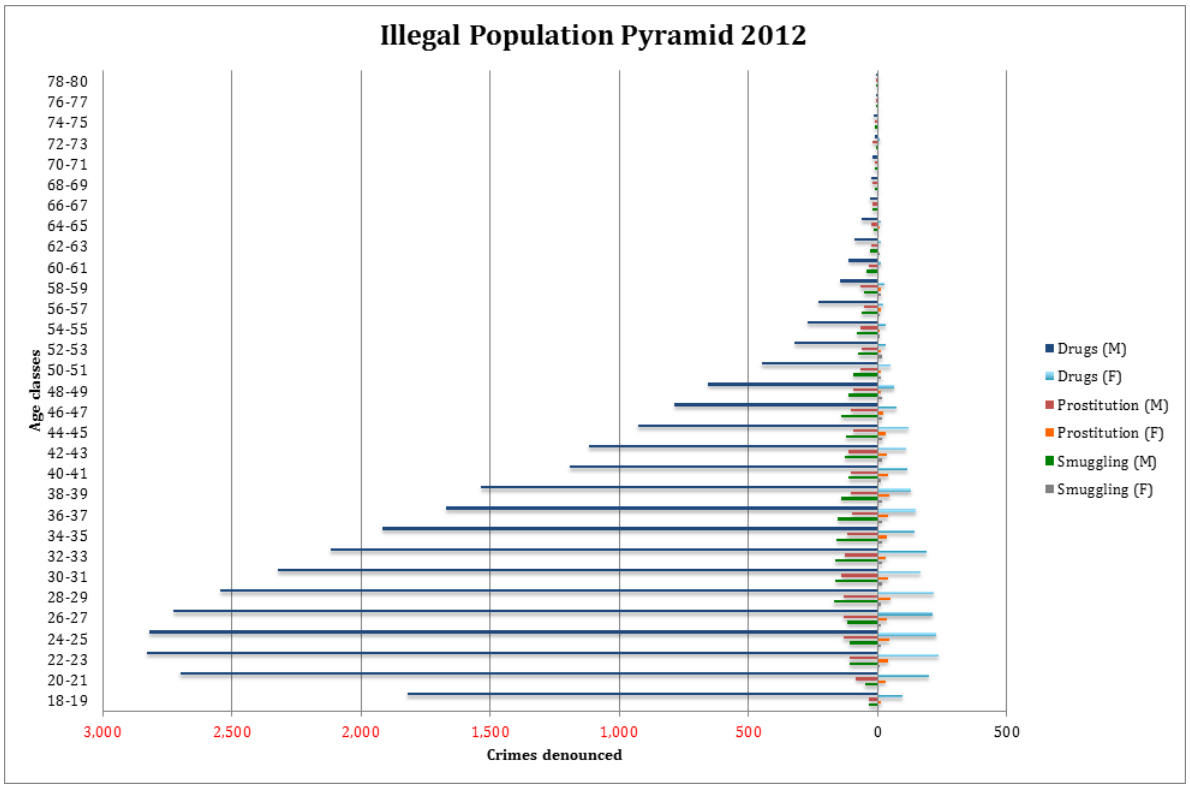

Figure 1: Illegal population age and gender pyramid in 2012

heterogeneity of the parameter. The covariate "age" goes to 18 until 80 the year when you reach the age of majority in Italy and from which you can go to prison and until the year that we considered possible to carry out a job, other data in the dataset . The covariate "gender" indicates obviously the sex of the criminals which is predominantly male. 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.

In the Figure 1 the pyramids describe the age and gender of the observed criminal population. The figure shows primarily that the denunciation of drug related crimes committed by men are far greater than the denunciation made for other crimes. Prostitution and smuggling related crimes data tell us also 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 denunciation for all the three type of crimes, but for drug trafficking it is more evident than in others.

The Table 3 reports the covariates for drugs related crimes in 2013: as already shown by the pyramids for 2012 in Figure 1, male criminals are far more than female, most criminals are concentrated in the age group under the age of 50. Italians are only twice than foreigners. Moreover, most of the criminals do not have denunciation about other types of crimes, even if the denunciation for other crimes are not few, the counts are finally almost divided in half between those who acted alone and those who acted in

Table 3: Counts for covariates for drug crimes in 2013

Covariate	f_1	f_2	Counts
Female	2276	275	2732
Male	27479	3833	33852
$\leq$30 years	15475	2123	18952
30-50 years	12556	1778	15555
$>$ 50 years	1724	207	2077
Italians	19431	2702	23917
Foreigners	10324	1406	12667
Not-involved in other crimes	23523	2974	28397
Involved in other crimes	6232	1134	8187
Act alone	17733	1390	19525
Act in association with other people	12022	2718	17059

Table 4: Models for drugs related crimes in 2013

Model	AIC	G^2 Test	N	C.I.
G+A+N+OC+As	24017.14	Accept remove A	183064	175734 - 190394
G+N+OC+As	24015.28	Accept remove N	183061	175732 - 190390
G+OC+As	24013.35	Reject remove none	183057	175728 - 190386
G+OC	24940.86	Reject remove OC	154025	149318 - 158732
G	25026.03	Reject remove G	151782	147263 - 156301
OC	24941.38	Reject remove OC	153937	149239 - 158635
Null	25028.90		151625	147122 - 156128
As	24009.35	Reject remove As	181460	174264 - 188656

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.

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". This is a result that we expected because, as we have seen from the descriptive analysis of the data, the crimes related to drugs are done more by men and the fact that they have done other types of crime and that they have made them in association to other people strengthens the thesis that the drug related crimes are the typical crimes done within national organization that deals with other crimes also.

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	183061	175732 - 190390	187518	176992 - 198044
G+OC+As	183057	175728 - 190386	186124	178628 - 193620
G+OC	154025	149318 - 158732	156687	151868 - 161506
Null	151625	147122 - 156128	154253	149642 - 158864

4.2 The linkage errors

As stated in section 3.2 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 working on legal activities, i.e. the labour force survey sample 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 criminal register. Comparing the results of linkage performed via the person identifiers with the results from the linkage based on soft attributes we asses the probability of being linked by chance in the criminal 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 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 creimes 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 relared crimes, according to the methodology illustrated in section 3.2. 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.

In Figure 2 the population size for drugs is estimated for the 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 avaliable covariance on Nationality.

Finally, Figure 3 shows the population size estimates for all the considered crimes, i.e. drug, prostitution and smuggling. It is worthwile to notice the unespected behavior of the estimate for drugs in 2009. The fall in 2009 is not obeserved in the other crimes and it

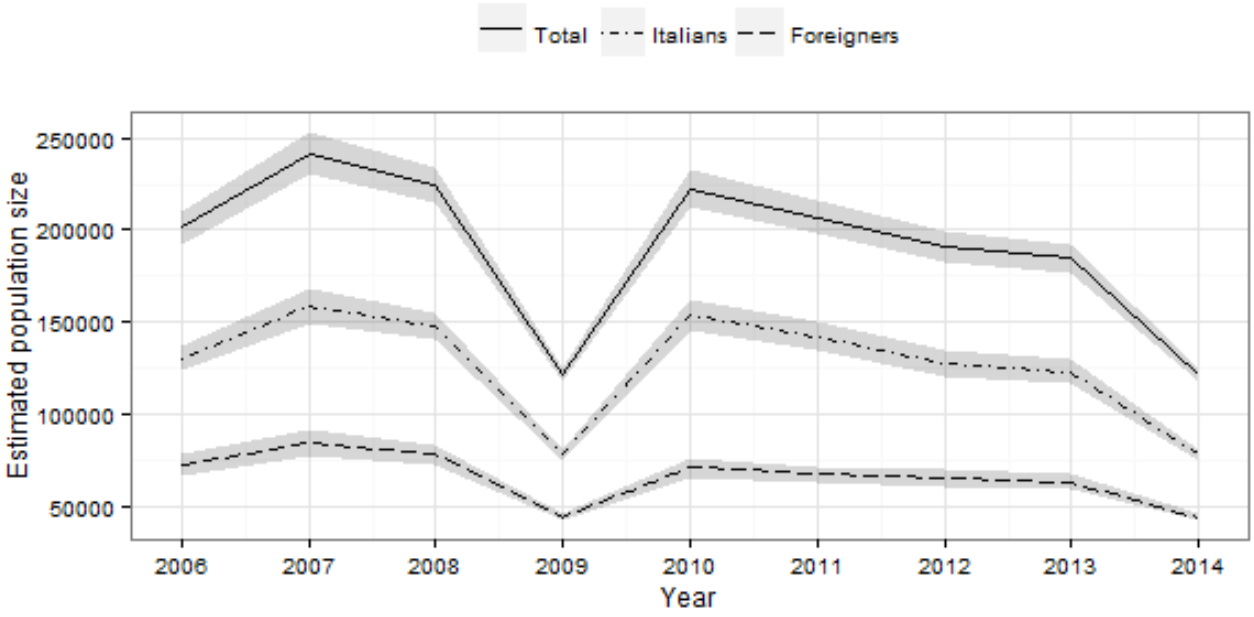

Figure 2: Estimated population size for drug by year. Subpopulations of Italians and Foreigners. Confidence Intervals in the coloured area.

is not justified by a fall in the number of proceedings. It is only justified by the different behavior of the covariate "Act in association" in 2009.

5 Concluding remarks and future works

The literature on estimating illegal populations with the Zeltermann estimator already exists but it is still poor. 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 criminal identifiers in the dataset.

This study provides an estimate of the population of illegal actors 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 Zeltermann estimator. Moreover, due to the lack of an exact identifier for criminals in the data, the Zelterman estimator needs an adjustment for the linkage error. The comparison with other data related to regular workers suggests to consider 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

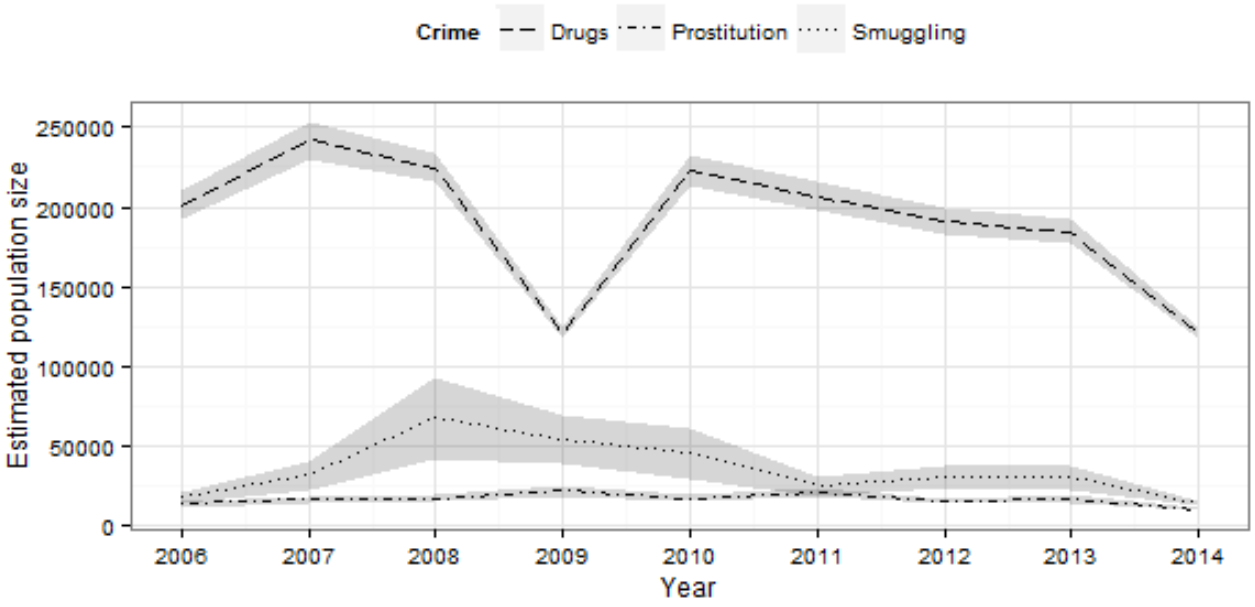

Figure 3: Estimated population size for drug, smuggling and prostitution by year. Confidence Intervals in the coloured area.

association, having denunciations about other crimes and gender (males are much more prone to commit crimes).

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, unless errors due to non-statistical nature of the collected data, it could be the first stepaiming at providing accurate estimates of the illegal population in Italy. Further analyses will be dedicated to calculate other populations of illegal actors such as those who commit corruption or who operate in the mafia organizations considering characteristics specific of the crimes.

References

[1] Blumstein, Alfred, ed. Criminal Careers and" Career Criminals,". Vol. 2. National Academies, 1986.

[2] Bouchard, Martin, and Pierre Tremblay. "Risks of arrest across drug markets: A capture-recapture analysis of hidden dealer and user populations."Journal of drug issues"35.4 (2005): 733-754.

[3] Collins M.F., Wilson R.M. 1990 Automobile theft: Estimating the size of the criminal population. Journal of Quantitative Criminology, 6 (4)

[4] Greene M.A., Stollmack S. 1981 Estimating the number of criminals. In Fox James A. (Ed.), Models in Quantitative Criminology. New York:

[5] Relazione annuale al Parlamento 2015 sullo stato delle tossicodipendenze in Italia

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

[7] 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: 122-129.

[8] Rossmo D. K., Routledge R. 1990 Estimating the size of criminal populations. Journal of Quantitative Criminology,

[9] Van Der Heijden, Peter GM, Maarten Cruyff, and Hans C. Van Houwelingen. "Estimating the size of a criminal population from police records using the truncated Poisson regression model." Statistica Neerlandica 57.3 (2003): 289-304.

[10] Van Der Heijden, P.G.M., Bustami, R., M. Cruyff, G. Engbersen and H. van Houwelingen (2003). Point and interval estimation of the truncated Poisson regression model. Statistical Modelling, 3, 305-322.

[11] 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, 595610.

[12] Zelterman D. (1988), Robust estimation in truncated discrete distributions with application to capturerecapture experiments, J. Statist. Plann. Inference 18 (1988) 225237