Summary

Estimation methods for sizing hidden populations, specifically the illegal migrant population

Introduction

In the *Hoofdlijnenakkoord* of the second Balkenende administration it was proposed that illegal residence by foreigners in the Netherlands will be counteracted more strongly than before. The illegally resident foreigners in the Netherlands form an example of a hidden population. In order to evaluate the effects of the intended policy it is therefore necessary to estimate the size of this kind of hidden populations. This report gives an overview of methods that can generally be used to make inferences about the size of hidden populations. More particularly the suitability of the methods is discussed to determine the size of populations of illegal foreigners and the sizes of subpopulations of these groups, for instance by nationality or by place of residence.

Formally, populations can be hidden for two reasons:

– No list exists on which all members of the population are listed, in other words, naturally there is no sampling frame. In the case of illegal foreigners, there are lists on which a number of them appear, for instance a medical registration, police files, schools pupil files etc., but there is no complete list on which every illegal foreigner appears.
– There is a list on which all members of the population are listed, but many more people appear on the list, and it is not a priori clear who does and who does not belong to the hidden population. In this case we speak of a 'hidden population' when it is difficult to determine who belongs to it. The question whether someone wants to change jobs is easily asked and answered, but medical questions or questions about socially undesirable behaviour or criminal behaviour are much more awkward.

The size of a hidden population often is a statistic which can only be determined with much effort. Therefore it is important to gain insight in all available alternatives to reach this goal. This report is mainly based on a search of the available literature; with respect to some issues, practical experience some authors with the subject matter has also been given a place in the text.
For the definition of an illegal foreigner, two points of view are possible: the economic and the demographic point of view. According to the economic point of view, a foreigner is illegal when he has a paid job without a valid working licence. In the demographic definition, the residence permit is central. Then, an illegal foreigner is someone who stays in the Netherlands

without a valid residence permit. It is irrelevant whether this person entered the country legally or illegally. This is the definition which is used is the major part of this report.
The report was written with a number of different goals in mind. In the first place it is a general overview of statistical methods which can be used for the estimation of the sizes of hidden populations. In this sense the report is meant to be a reference book. This overview is given in chapter 2 in a number of separate sections. Not only has been attempted to explain the technical working of the methods, but also to present the reader with an impression of how the methods relate to the state of the art of statistical methodology. He who, after reading such a section, mentions the method in a presentation on an international conference will not be surprised by totally new insights that appear to be part of accepted science. In the second place the report aims to give an insight into suitable methods for counting illegal foreigners. Not every method which is used for counting hidden populations is suitable for this purpose. Every method starts with a certain idealised image of reality. The actual reality is often much more complex and messy. For this reason, in chapter 3 an overview is given of a number of practical examples both in the Netherlands and abroad. It will appear that some methods are useful, but they always have to be supplemented with specific adaptations. In describing the practical applications, the third goal is to give the reader a general idea of the problems that appear when illegal foreigners are counted in practice. Finally, and this is the goal of chapter 4, the report gives information and support that can be used for choosing a specific method for executing a longitudinal estimation procedure of the size of the population of illegal foreigners.

The methods

The methods which are discussed in this report vary considerably with respect to procedure, assumptions and required input. Some of them have already been used for the estimation of the size of illegal populations, others have been used for the estimation of the size of other hidden populations but not for illegal populations; there are also methods of which the use is conceivable, but for which practical experience is still lacking.

Capture-recapture methods

Capture-recapture methods form a class of estimation techniques for which there exists a considerable body of experience, especially in the Netherlands. These methods are ways of doing survey research without sampling frame. Sample elements are literally ‘caught’, but also released again. By drawing different samples and determining how many elements appear in more than one sample, the size of the population from which they are drawn can

be estimated. The name 'capture-recapture' originates from biology, where this method is used for the estimation of population sizes of animal species. It is attempted to catch the animals repeatedly. In biology it is customary to draw the samples at different moments in time. Literally, the animals are caught, and later recaught. But it is also conceivable that two catchers catch and release the animals independently from each other. In this way the draw samples which are constructed more or less simultaneously. This metaphor usually applies to examples in epidemiology and government statistics. In that case, existing records in different registrations are used, in which the sample elements are 'caught'. After linkage, the numbers of persons from the target population that appear in one or more of the registrations can be determined. Under certain conditions the number of persons that do not appear in any of the records in the registrations can be estimated. Together, the number of linked persons and the estimated number of absent persons constitute the estimation of the size of the hidden population.

The use of linked registrations

In order to apply capture-recapture methods on the basis of linked registrations to the population of illegal foreigners, a number of conditions have to be met. These conditions are the following:

– closed population: during the period to which the estimates relate, the population must remain constant
– no linking errors: the persons are correctly and uniquely identified
– independent inclusion in registrations: the probability to appear in all registrations is the product of the probabilities to appear in each of the registrations separately.

These conditions are not always met. In many respects it is known how violations of the conditions influences the estimates; more in particular, it is known whether they lead to an upward or a downward bias. Finally, there is a fourth condition, which is more a wish than a requirement.

– the probability to appear in a registration is identical for all members of the hidden population

When this condition is not met, statistical methods are available to handle these differences in probability. They have to be included in the model and estimated, which makes the statistical calculations of the sizes of the hidden populations more complicated, but not impossible.

Repeated capture

When samples are captured repeatedly (illegal foreigners are apprehended repeatedly), the estimation procedure of the size of the hidden population is based on one single data set, in which illegal foreigners may appear

more than once. For the number of appearances a theoretical distribution, the Poisson distribution, is used, which indicates the probability that a person is caught 0, 1, 2 times etc. This distribution depends on a parameter, which is estimated on the basis of the data, after which the probability can be determined that an individual is never caught. This, in turn, leads to estimation of the size of the total population. The most crucial assumption is that this distribution can actually be used. It is assumed that the probability of recapture, after a first capture, remains unchanged. An assumption which, like in the case of linked registrations, can be avoided with the use of complex statistical calculations, is that every illegal foreigner has the same Poisson-parameter. When this assumption is not met, the differences between the Poisson-parameter can be included in the model and explicitly be estimated.

Snowball sampling

Snowball samples are another example of samples without sampling frame. The first step in a snowball sample is a small initial sample. Next, the drawing of new elements is left to the respondents already in the sample. There are two variants of this kind of samples: (i) the 'classic' snowball sample, where the respondents mention to the fieldwork organisation the names of new potential respondents and (ii) *respondent driven sampling*, where the respondents themselves approach new potential respondents with the request to participate to the survey. This last approach has the advantage that nobody discloses the identity of another person. In the framework of research of illegal foreigners, this is a considerable advantage, as from previous research it appeared that there is much fear to participate. It even is questionable whether this fear can be removed by the use of respondent driven sampling. Presently, this question cannot be answered, because the method has never been applied within this context. The construction of estimators on the basis of the outcomes of snowball samples is statistically complex. For the classical approach, handy formulas are available only in the case of an initial sample and one single wave of new respondents. In the case that the number of illegal foreigners has to be calculated as a percentage of a larger population (e.g. all foreigners), a more general method is available. In this method one has to account for the fact that people with a large social network are overrepresented. This makes the estimation process complex.

Detection Controlled Estimation

Detection Controlled Estimation (DCE) is an estimation technique to simultaneously estimate the characteristics of offenders and their inspectors. This may concern persons, e.g. students who may commit fraud during their exams, or companies that violate environmental regulations. In both cases there is an inspector. In the case of exams, this is the supervisor; in the case

of companies this is the investigator. In both examples, the sampling units are pairs: an inspector is assigned to every potential offender. An offense is registered only when it is made by a potential offender and it is discovered by the inspector. Undiscovered offenses are not registered. Estimation of the size of a hidden population therefore amounts to estimate the number of undiscovered offenses. The probability of making an offense is denoted by p_1; the probability of discovery of an offense is denoted by p_2. Then the probability of a discovered offense is equal to p_1p_2. The probability that this is not the case is equal to $1\text{-}p_1p_2$. In this last case it is unknown whether an offense occurred; is is only known that no offense has been discovered. The probabilities of offenses and their discoveries are linked by the technique to the characteristics of the offenders and the inspectors. Once this relationship has been estimated, it is also possible to make inferences about the number of undetected offenses. In this respect, 'being an illegal foreigner' may be interpreted as an offense. The inspector may be a police officer, but also a respondent in a survey, who reports on his knowledge of illegal foreigners in his neighborhood.
DCE is a very flexible econometrical technique, which allows a wide variety of refinements in the formulation of the model. For instance, there may be foreknowledge at the inspectors at the strategic level (they know the offenders characteristics), at the tactical level (they see the actual offense coming) and the possibility may be included that the offenders self-report their offenses. These adaptations come, however, with a price. Estimation with these advanced models require high level programming, an therefore a long production time to execute the estimation process.

Estimation based on postal code areas

The Netherlands are partitoned into a refined network of postal codes. Of these postal codes, characteristics are known at the most detailed level (6 digits). Of course, also the national distributions of these characteristics are known. This makes it possible to use the database of postal codes as a sampling frame. For hidden populations that are concentrated in postal code areas with specific characteristics, this allows for the possibility to use the postal code characteristics for efficient estimation of their sizes at the national level. Obvious examples are the homeless, drug addicts, prostitutes and also the illegal foreigners, as in police administrations the postal code of their place of residence is included. From 1998 until 2003, 11517 postal code areas illegal foreigners who were apprehended were found to live in each month. In 2003, more than 100 of such postal code areas are added. Analysis of data from Experian, a firm which sells postal code area segmentations, showed that these areas are strongly linked to economic indicators like purchasing power and amount of debts of the inhabitants. These data and regularities make it possible to draw a well stratified sample of postal code areas. However, after that follows the most uncertain part : the estimation

of the number of illegal foreigners within each postal code area. To this end, a number of approaches are possible: (i) determination by the police, (ii) observation by trained experts, (iii) face to face interviews by well trained interviewers within the postal code areas, (iv) a survey using a large internet panel (market research firm TNS NIPO appeared to have a sufficient number of respondents which live in postal code areas where illegal foreigners were apprehended). It is, however, unknown whether this results into valid estimates per postal code area. An encouraging indication is the fact that in the U.S.A. the number of illegal foreigners is determined on the basis of face to face interviews; the levels of underreporting are considered to be acceptable. It is, however, not clear that these findings also apply to the Dutch situation.

Randomized response

Randomized response is a method designed for the measurement of sensitive opinions, attitudes and behaviour, like violations of the law, use of drugs and alcohol and sexuality. Respondents often perceive questions about these types of subjects as threatening; this leads to an inclination to give socially desirable answers. Randomized response is a method that is specifically designed to minimize the threatening nature of these questions and in this way increase the validity of the answers. By the randomized response procedure, the answers to sensitive questions are partly determined by chance. This can be achieved in various ways. For instance, in the so called 'forced response' method the respondent throws two dice and then answers the question on the basis of the number of dots. On 2, 3 and 4 dots he is obliged to say yes, on 11 of 12 dots he is obliged to say no. On all other numbers the respondent answers truthfully. In this way, nobody knows whether an obligatory answer has been given, and the privacy of the respondent is guaranteed. On the basis of statistical analysis of the given answers, the distribution of the true answers can be deduced. Randomized response is hard to conceive in a direct survey among potentially illegal foreigners. ('are you illegal?', yes/no), but it is a method that can be used for obtaining information about illegal foreigners, e.g. among employers who may employ them illegally.

De Delphi method

The Delphi method is not a statistical technique, but a procedure that is based on combining the judgments of experts. The basic procedures of the Delphi method are the following. (1) A panel is formed of experts who all have their individual view on the subject matter. (2) The experts have no direct contact. They learn about each other's opinions through a moderator, who distributes the input anonymously. (3) In the first round, the

experts state their opinions on the problem area, in this case the size of the population of illegal foreigners. (4) In the second round, they are confronted with each other's opinion; they adjust their opinions and/or motivate their unchanged or adjusted opinions. (5) In the following round, this procedure is repeated. Ideally this process continues until the opinions converge (and there is one common estimate of the size of the population) or, alternatively, until the different opinions stop changing. The purpose of the method is to make use of the joint knowlegde of experts, without disturbing factors that may play a role in personal discussions like conflicts and the domination by the most verbally gifted persons or strongest egos.
Although the method is used on a very large scale and for a large variety of problems, it is not undisputed. As a rule, no use is made of a clear empirical line of reasoning. Still, the procedure has been statistically investigated. Not surprising, but still useful is the overwhelming evidence that consensus after the first round usually increases. Apparently, the experts tend to listen to each other and learn from each other. It also appears from the comparison with already known figures or short term forecasts that the results of a Delphi procedure at the end are more accurate than in the first round. Moreover, Delphi appears to give more accurate results than comparable group processes. In other words: where a scientific evaluation is possible, the method appears to be superior to comparable alternatives. Still, the use of the Delphi method as a basis for a monitor of sizes of hidden populations seems to be problematic. Experts without statistical training will have have to extrapolate figures from their local experience to the national level. Moreover, it will be hard to keep the line of reasoning constant over the years, which is a threat to the comparability of the figures through time.

Other methods

Chapter 2 concludes with some short descriptions of a number of other methods. The *demographic method* departs from a given point in time when the size of the hidden population is known. On the basis of demographic variables like fertility, mortality and emigration it is calculated how the expected size of a cohort of illegal foreigners develops through time. The *multiplier methode* is a way to use known proportions to estimate an unknown figure. It is, for instance, possible to make use of the fact that one out of 10 heroin users will die of an overdose sooner or later. The number of victims of an overdose at a certain place or time may be used for the estimation of the total number of heroin users: when there are 50 victims of an overdose, the total number of heroin users is estimated to be 500. With the *residual method* the number of illegal foreigners is calculated as the result of a subtraction. The usual way of doing this is to make an estimate of the total number of foreigners, an to subtract the number if legal foreigners. Finally, the *three cards method* is a new alternative to randomized response for asking sensitive questions. By ingenious variation of question texts on

three different cards, one of which is given to the respondent, an estimation can be made of the percentage that falls into a sensitive category.

Application of the methods in practice

Whereas the overview of the different methods forms a reasonably clear list of possibilities for the estimation of numbers of illegal foreigners, the description of the international practice, in particular, is a messy set of partly succeeded attempts, lacking information and ad hoc adaptations without a systematic approach. For instance in *Marocco* the number of emigrants was determined on the basis of a demographic analysis. By subtracting the number of legal emigrants, the illegal emigration could be deducted. So far, the calculations were carried out in a plausible manner. However, next this number was allocated to the different European countries proportional to legal emigration. For this calculation there was no empirical basis. An attempt to conduct the same analysis for *Tunesia* failed due to the bad quality of the Tunesia census. In *Portugal* immigration data were compared with census data from 1981. The number of illegal foreigners was estimated as the difference between the number of foreigners in the census and the legal immigrants. In 1991 there appeared to be confusion about the difference between place of birth and place of residence, which, in retrospect, rendered the estimate worthless. In *Belgium* it was established in 1990 that in the neighbouring countries the illegal foreign population was estimated to be 10% of the total foreign population. This figure was copied, but never officially published. In the *Czech* republic a study was made of loopholes in the law. Next, a panel of experts (researchers, immigration officials) judged to what degree these loopholes were used by illegal foreigners and employers. Unfortunately, the results of this study were never published.
These examples (many more are described in chapter 3) make clear that attempts to count illegal foreigners only lead to plausible results when considerable investments are made in resources and quality. Two countries that are good examples in this respect are the U.S.A. and South Africa. The country which provides a good application of the *residual method* is the U.S.A. Three large surveys can be used: the census, the supplementary ACE (Accuracy and Coverage Evaluation) and the CPS (Current Population Survey). These data are combined with those of the immigration service IND. The US Bureau of the Census started in 1980 counting the illegal immigrants, based on the census. On the basis on the questions in the census, in 1980 it was determined who was foreign-born and who was legal or illegal.

The general framework for counting the number of illegal foreigners was a formula by which the number of foreign born in the U.S.A. was calculated:

$$\text{Foreign born} = L - (M+E) + T + R$$

L the number of legal foreigners
M the total mortality among foreigners
E the total emigration of foreigners
T the number of temporary resident foreigners
R the number of unauthorized foreigners; this was partitioned into the number of foreigners who still are in an asylum procedure and the real illegals.

The components were estimated on the basis of different sources, which were analyzed by different teams.

In *South-Africa*, estimates were based on flow data (immigration and emigration) of three different groups: (1) legal entry, illegal stay, (2) illegal entry, legal stay and (3) illegal entry, illegal stay. Besides, a large survey was carried out in the countries of origin to investigate the nature of the illegal stay. A surprising result was that as a rule the duration of the stay of illegal foreigners in South-Africa was short (several months). After that, they returned to the country of origin. For the investigators this was a reason to put the problem of illegal foreigners in Sout Africa somewhat in perspective.

Research in the Netherlands

Within the Netherlands there is a rich experience with the estimation of numbers of illegal foreigners; especially in the field of more advanced statistical methods much work has been done in the Netherlands. There is a limited number of publications in which emphasis is placed on illegal employees. As early as 1994 there were publications in which estimates were presented based on impressions of the *Loon Technische Dienst* and the *Belastingdienst* and based on interviews with employers. Recently, in 2005, in a report by *Regioplan* an account was given of the compliance to a law which regulates work by foreigners (*Wet Arbeid Vreemdelingen*). In a survey among employers the *randomized response* method was applied.
In 1995 the number of illegal foreigners was estimated in the four big cities in the Netherlands, based on an estimate for Rotterdam. The estimate for Rotterdam was constructed by making an estimation of the number of criminal illegal foreigners in Rotterdam, and combining this with the ratio criminal/non-criminal illegal foreigners obtained in other interviews. This combination can be seen as an application of the *multiplier* method.
The majority of the applied methodological publications deals with capture-recapture methods. There is experience with both main variants of the method: linked registrations and repeated capture. On the basis of repeated capture, estimates of the size of the illegal population in the period 2000-2003 are made on the basis of the regional *Vreemdelingen Administratie Systeem* (VAS, Foreigner Administration System), which at the end of march 2005 has been replaced by the national *Politie Suite Handhaving-*

Vreemdelingen (PSH-V, Police Suite Enforcement-Foreigners). In principle, all apprehended illegal foreigners are registered in this system. In doing so, the model had to be adapted for illegal foreigners who came from Eastern Europe (usually they stay for a short period, so they are not a closed population) and illegal foreigners who were effectively expelled from the Netherlands (in that case the Poisson model did not apply). The fact that the method leads to relatively large confidence intervals appeared to be a serious practical problem.
A second relevant estimation method that was used in the Netherlands is the capture-recapture method on the basis of *linked registrations*. This is applied to estimate the number of Antillians who are not registered in the *Gemeentelijke Basis Administratie* (GBA, the Dutch population register). Given the special relationship between the Antilles and the Netherlands, formally there are no 'illegal Antillians'. More in general, there often is a large overlap between being illegal and not being registered in the GBA. The data were combined with the *HerkennингsDienst Systeem* (HKS, recognition service system) of the police. Execution of the estimation method appeared to be relatively simple, with acceptable confidence intervals.

How to continue?

In chapter 4 the discussed methods are systematically compared with respect to their usefulness in the near future. Most of the methods are not qualified, because they require a considerable amount of preliminary research. The answer to the question whether they can really be used is uncertain. Incidentally, on the basis of this report it is clear how such preliminary research has to be conducted. Capture-recapture methods can be used immediately; they have proved to work, although their underlying assumptions are not met perfectly. The method is cheap, because the data are already available. However, the size of the confidence intervals with the repeated capture method is a problem. When linked registrations are used, it may be difficult to base the estimations on data that are sufficiently recent for policy purposes.